# Bases génétiques de l'adaptation du moustique tigre Aedes albopictus à de nouveaux environnements : une approche sans à priori reposant sur les éléments transposables

Clement Goubert

N° d'ordre : 276-2015 Année 2015

Thèse de l'université de Lyon
délivrée par L'Université Claude Bernard Lyon 1
école doctorale : Evolution, Ecosystèmes, Microbiologie et Modélisation

pour l'obtention du
**Diplôme de doctorat en Biologie**
(arrêté du 7 août 2006)

soutenue publiquement le
4 décembre 2015

par
Clément Goubert

# Bases génétiques de l'adaptation du moustique tigre *Aedes albopictus* à de nouveaux environnements : une approche sans a priori reposant sur les éléments transposables

Directeur de thèse : Monsieur Matthieu Boulesteix

Co-directrice de thèse : Madame Cristina Vieira

| | | |
|---|---|---|
| Jury : | Monsieur Pierre Capy | Rapporteur |
| | Monsieur Frédéric Simard | Rapporteur |
| | Madame Josefa González | Examinatrice |
| | Monsieur Denis Bourguet | Examinateur |
| | Monsieur Frédéric Fleury | Examinateur |
| | Monsieur Matthieu Boulesteix | Directeur de thèse |
| | Madame Cristina Vieira | Co-directrice de thèse |

### *UNIVERSITE CLAUDE BERNARD - LYON 1*

| | |
|---|---|
| **Président de l'Université** | **M. François-Noël GILLY** |
| Vice-président du Conseil d'Administration | M. le Professeur Hamda BEN HADID |
| Vice-président,du Conseil des Etudes et de la Vie Universitaire | M. le Professeur Philippe LALLE |
| Vice-président du Conseil Scientifique | M. le Professeur Germain GILLET |
| Directeur Général des Services | M. Alain HELLEU |

### *COMPOSANTES SANTE*

| | |
|---|---|
| Faculté de Médecine Lyon Est – Claude Bernard | Directeur : M. le Professeur J. ETIENNE |
| Faculté de Médecine et de Maïeutique Lyon Sud – Charles Mérieux | Directeur : Mme la Professeure C. BURILLON |
| Faculté d'Odontologie | Directeur : M. le Professeur D. BOURGEOIS |
| Institut des Sciences Pharmaceutiques et Biologiques | Directeur : Mme la Professeure C. VINCIGUERRA |
| Institut des Sciences et Techniques de la Réadaptation | Directeur : M. le Professeur Y. MATILLON |
| Département de formation et Centre de Recherche en Biologie Humaine | Directeur : Mme. la Professeure A-M. SCHOTT |

### *COMPOSANTES ET DEPARTEMENTS DE SCIENCES ET TECHNOLOGIE*

| | |
|---|---|
| Faculté des Sciences et Technologies | Directeur : M. F. DE MARCHI |
| Département Biologie | Directeur : M. le Professeur F. FLEURY |
| Département Chimie Biochimie | Directeur : Mme Caroline FELIX |
| Département GEP | Directeur : M. Hassan HAMMOURI |
| Département Informatique | Directeur : M. le Professeur S. AKKOUCHE |
| Département Mathématiques | Directeur : M. le Professeur Georges TOMANOV |
| Département Mécanique | Directeur : M. le Professeur H. BEN HADID |
| Département Physique | Directeur : M. Jean-Claude PLENET |
| UFR Sciences et Techniques des Activités Physiques et Sportives | Directeur : M. Y.VANPOULLE |
| Observatoire des Sciences de l'Univers de Lyon | Directeur : M. B. GUIDERDONI |
| Polytech Lyon | Directeur : M. P. FOURNIER |
| Ecole Supérieure de Chimie Physique Electronique | Directeur : M. G. PIGNAULT |
| Institut Universitaire de Technologie de Lyon 1 | Directeur : M. le Professeur C. VITON |
| Ecole Supérieure du Professorat et de l'Education | Directeur : M. le Professeur A. MOUGNIOTTE |
| Institut de Science Financière et d'Assurances | Directeur : M. N. LEBOISNE |

**Résumé**

Le moustigre tigre *Aedes albopictus*, un des vecteurs de la Dengue et du Chikungunya, est une espèce invasive qui a colonisé le monde entier en 30 ans à partir de son berceau asiatique. Les éventuelles bases génétiques de ce succès sont inconnues. Afin d'étudier l'ampleur de la différenciation génétique entre populations asiatiques et européennes et la part prise par la sélection naturelle dans cette différenciation, nous avons développé de nouveaux marqueurs génétiques reposant sur le polymorphisme d'insertion des éléments transposables. Pour cela, nous avons dans un premier temps conçu un outil bioinformatique –dnaPipeTE— nous permettant de dresser le portrait de la fraction répétée du génome d'Ae. albopictus à partir d'une faible proportion des lectures brutes issues d'un projet de séquençage en cours. Le polymorphisme d'insertion de cinq des familles d'ET décrites a ensuite été étudié par la technique de transposon display couplée à du séquençage Illumina, chez 140 individus issus de trois populations vietnamiennes et cinq populations européennes. L'immense majorité des 128 000 marqueurs analysés montre une différenciation génétique très faible entre Europe et Asie. Nous avons néanmoins pu mettre en évidence un centaine d'insertions ayant des fréquences extrêmement différentes entre ces continents. La majorité d'entre elles ségrège à forte fréquence en Europe, suggérant une adaptation du moustique à son environnement tempéré.

**Abstract**

The Asian tiger mosquito, one of the main vectors of Dengue and Chikungunya, is an invasive species that colonized the world during the last 30 years from its cradle in Asia. Whether this success has an underlying genetic basis remains to be investigated. In order to study the extent of the genetic differentiation between Asian and European populations and the contribution of natural selection to this differentiation, we developed new genetic markers based on transposable elements insertion polymorphism. We first conceived a bioinformatic pipeline –dnaPipeTE— that allowed to grasp a comprehensive picture of the repetitive fraction of the Tiger's genome through the analysis of a low proportion of raw reads from a ongoing sequencing project. The insertion polymorphism of five transposable element families was then studied by Illumina based transposon display, in 140 individuals from three Vietnamese populations and five European populations. The vast majority of the 128,000 markers showed a very low genetic differentiation between Europe and Asia. However 92 of them displayed extreme frequency differences between the continents. The majority of them segregate at high frequencies in Europe, a pattern suggestive of adaptive evolution towards temperate environments.

# Remerciements

Mes premiers remerciements sont pour mon directeur de thèse, Matthieu Boulesteix. Je suis très fier d'avoir été ton premier doctorant, et je tiens à te remercier pour la confiance que tu m'as accordée dès le début de cette thèse. Merci de l'attention que tu as portée au bon déroulement de ce projet, et aussi d'avoir toujours veillé à ce que je me sente à l'aise durant ces trois ans. Merci pour ta disponibilité, tes conseils et tes commentaires, toujours précis et constructifs. Et même si tu dis parfois "ergoter", ce n'est jamais pour rien, et je sais que cela m'a beaucoup apporté !

Un grand merci aussi à ma co-directrice ce thèse, Cristina Vieira. Tu as énormément contribué à distiller la motivation et l'énergie nécessaires pour mener à bien cette expédition. Je tiens en particulier à te remercier pour ta confiance et ton enthousiasme pour l'ensemble des projets menés et initiés.

Matthieu, Cristina, j'ai eu beaucoup de chance de réaliser ma thèse avec vous. Merci à tous les deux pour la liberté et l'autonomie, la pression justement dosée et l'excellente ambiance, au travail et en dehors.

Je tiens à remercier les membres du jury qui ont accepté de lire et évaluer ce travail : les rapporteurs, messieurs Frédéric Simard et Pierre Capy ainsi que les examinateurs madame Josefa Gonzáles, messieurs Denis Bourguet et Frédéric Fleury.

Je remercie les membres de mes comités de thèse : mon tuteur, Christophe Douady ainsi que Josefa Gonzáles, Patrick Mavingui et Aurélie Bonin, qui ont par leur intérêt et leurs conseils contribué à la bonne marche de ces travaux.

Cette thèse n'aurait simplement pas été possible sans la collaboration de Patrick Mavingui, Claire Valiente Moro et Guillaume Minard qui ont mis à notre disposition leurs échantillons précieux, les ressources génomiques ainsi que toute leur expérience concernant notre modèle commun. Merci Guillaume pour ton aide et ta réactivité sans faille, j'espère que nous aurons l'occasion de travailler à nouveau ensemble !

Je dois aussi beaucoup à Laurent Modolo, formidable colocataire de bureau et bien plus encore. Tes précieux conseils et tes questions où l'on dit d'abord "ha oui tiens..." et auxquelles on réfléchit toujours des heures plus tard ont joué un rôle très important dans cette thèse. Merci pour ton aide, je me suis grâce à toi initié avec plaisir aux joies geek des lignes de codes et des téléphones intelligents, et puis surtout : on s'est bien marré !

Je souhaite aussi remercier Hélène Henri, véritable Shiva de la biomol', tu as eu assez de bras pour t'occuper de mon cas avec attention, à veiller au grain, et avoir permis, la bonne humeur en prime, que notre transposon display fonctionne au poil !

Un grand merci à Manon Vigneron et Matthieu Cortes, mes stagiaires de choc. Vous avez assuré en faisant chauffer à blanc les thermocycleurs du LBBE à la gloire de cette thèse !

Il y aurait encore un affreux bug dans dnaPipeTE sans Valèria Romero Soriano, mais ça serait résumer ta contribution à peu de choses (notamment, désolé pour les piqûres de moustiques), alors tout simplement *gràcies*.

Merci à tous les transposons, ceux de passage et ceux dans la place : Hélène Lopez, Marie Fablet, Annabelle Haudry, Emmanuelle Lerat, Nelly Burlet, Virginie Braman, Elias Gutierrez, Judit Salces, Bianca Menezes, Emanuel Vilafan, Nicolas Bargues, Thérèse Callet, Gabriel Krazovec, Camille Simonet,... Merci à tous pour la très bonne ambiance au quotidien, votre disponibilité et tous vos conseils. Merci aussi à Sonia Martinez et Nicole Lara, que j'ai parfois un peu envahies avec nos activités extra-recherche !

Je voudrais aussi remercier Abdelaziz Heddi et toute son équipe pour m'avoir laissé jouer avec les charançons !

Quelle aventure au labo avec Christophe Plantamp (le premier des doctorants que j'ai rencontré !), merci de

m'avoir toujours laissé débarquer dans ton bureau à l'improviste pour papoter, réfléchir, et surtout : préparer l'Happy Hour !!!

Merci à Gabriel Terraz, membre fondateur du groupe informel "sécurité aérienne et génétique des populations en bash"

Parce que la biologie c'est aussi de la silice, un grand merci à Stéphane Delmotte et Bruno Spataro d'avoir si bien veillé sur l'armée des pbil's.

Merci en particulier Stéphane de t'être battu avec ardeur pour installer RepeatExplorer et de nous avoir convertis à cette formidable science à base de levures et de houblon.

Merci à Nathalie Arbasetti, Laetitia Catouaria, Odile Mulet-Marquis ainsi qu'Aline Maitrias pour votre support précieux (et votre gestion de mon catastrophisme organisationnel).

Je voudrais aussi remercier notre directrice d'unité, Dominique Mouchiroud pour son accueil chaleureux au sein du LBBE.

Un grand merci aux Docteurs Claire Schirmer et Romain Pierron qui se sont appliqués à relire ce manuscrit et le déminer de ses fautes d'orthographe.

La suite de ces remerciement s'adresse à tous ceux que je n'ai pas encore cités, et qui sont eux aussi à l'origine des très bon moments de ces trois années :

Aux membres du LBBE et fidèles de l'Happy Hour : Thomas, Murray, Simon, Rémi, Jeff, Magali, Héloïse, Fanny, Marie C., Laurent J. (*You wanna fight ?*), Thibault, Laurence, Patricia, Christelle, Adrian, Will, Wandrile, Adil, Sylvia, Gladys, Julien V., Natacha, Sylvain, Fabrice, Philippe.

Aux champions du monde de boulette en papier avec règles officielles : Olivier, Julien C. et Dave.

À l'ambiance lyonnaise : Eugénie, Les Modolettes, Robin, Jibé, Vincent et les ~~Genepi's brothers~~, ~~Minus one~~, Discount Jesus ? ? ? Merci aussi à Loïc, Maxou, Danaïe, Pili et Aurélie pour les guet-apens à la colloque !

À la côte ouest (la mer, le ski et le Cognac) : Claire, Flo, Delphine, Axel, Nono, JB, malgré la distance votre amitié m'est très précieuse !

Car le cheval c'est trop génial : merci à Pierre, Ewen, Nicolas, Antoine, Adrien et tous les rennais, une pensée aussi pour Céline, Mélanie, Sonia, Tiphaine, on en a fait du chemin depuis les staphylins !

Enfin, mes plus plates excuses à celles et ceux que j'aurais oubliés ici...

# Table des matières

*Pour mes parents Sylvie et Philippe,*
*pour ma soeur Léa*

# Introduction Générale

*"Harry, I have no idea where this will lead us, but I have a definite feeling it will be a place both wonderful and strange."*

– Dale Cooper, *Twin Peaks* 1991

# Avant propos

Les espèces invasives ont l'intérêt de nous permettre d'observer l'évolution en action, et en particulier d'évaluer l'importance du rôle joué par la sélection naturelle lorsque des populations se trouvent confrontées à un nouvel environnement. Au cours de cette thèse, nous avons cherché à savoir si des évènements d'évolution adaptative peuvent avoir contribués au succès invasif du moustique tigre *Aedes albopictus*.

Afin d'introduire les différents concepts abordés au cours de ces travaux, la première partie de ce chapitre est dédiée à la génétique de l'adaptation. Après de brèves définitions, nous nous intéresserons à la signature laissée par la sélection naturelle sur les génomes et évoquerons les différents tests et modèles permettant de la mettre en évidence. Nous aborderons aussi les particularités liées aux espèces invasives, l'influence de la colonisation sur la génétique des populations et la manière dont l'adaptation peut permettre l'adequation de ces espèces à leurs nouveaux environnements.

La recherche de nouveaux marqueurs génétiques chez *Ae. albopictus* nous a amené à étudier ses éléments transposables, une composante importante du génome chez cette espèce. La seconde partie de cette introduction est donc consacrée au éléments mobiles du génome, l'accent étant mis sûr la manière dont ils peuvent être utilisés en tant que marqueurs génétiques.

Enfin nous présenterons le modèle *Aedes albopictus*, ses caractéristiques écologiques et leurs influences quant à son statut actuel d'espèce invasive, avant de décrire notre approche concernant la recherche d'indices génomique de son adaptation à de nouveaux environnements.

# 1 La génétique de l'adaptation

## 1.1 Définitions

### Le processus adaptatif

En biologie évolutive, l'adaptation est un processus permettant l'ajustement des organismes vivants à leur leurs environnements biotique et abiotique au moyen de la sélection naturelle. Les changements liés au processus adaptatif ne pourront ainsi pas être observés à l'échelle de l'individu, mais ne le seront qu'en intégrant sa descendance, victorieuse du "combat pour l'existence" décrit par Darwin (1859) dans *L'Origine des espèces*. En invoquant ainsi l'action de la sélection naturelle, l'adaptation correspond à l'ensemble des modifications héréditaires qui de part leur influence sur le phénotype, ont conduit à l'augmentation de la valeur sélective (ou *fitness*) des organismes d'une population, dans l'environnement où ils se trouvent (Templeton 2006). L'adaptation a pour effet de sélectionner un phénotype que l'on pourra résumer à un ou plusieurs traits mesurables, par exemple la couleur, la taille, la forme ou même la présence d'un organe ou encore une fonction métabolique. Ce processus fait donc parti des forces évolutives à l'origine de la diversité des formes du vivant observable entre des environnements différents.

Parce qu'elle implique des variations au niveau du support de l'hérédité (*i.e.* la molécule d'ADN [1]), l'adaptation se distingue de la plasticité phénotypique, qui correspond au développement d'un individu vers un phénotype en réponse aux facteurs de l'environnement, mais sans modification de son génotype [2]. Autrement dit, pour opérer, l'adaptation a besoin de variabilité génétique.

L'adaptation correspond à une forme particulière de sélection, qualifiée de directionnelle : le changement d'environnement sélectionne des phénotypes qui transmettront une partie de la variance génétique et conduit l'établissement de nouvelles valeurs moyennes pour les traits considérés. A l'échelle du génome, on s'attend à observer des variations nucléotidiques, soit au cours de l'adaptation (dans le temps) soit entre individus adaptés à des environnements distincts (dans l'espace) au niveau des gènes ou des régions impliquées dans la réalisation des traits adaptatifs. L'effet d'une mutation favorable sur la valeur sélective, *via* une modification du ou des traits considérés, peut varier : on parle alors de "taille d'effet" (*effect size*) associée à un allèle (Orr 2005). On distingue d'une part les allèles à forts effets, portés par un nombre restreint de locus mais influençant fortement la *fitness*, comme ceux localisés sur les gènes *Eda* ou *pitx1* de l'épinoche, directement responsables de la réduction des plaques osseuses lors de l'adaptation à l'eau

1. Nous ne traiterons pas ici des modifications épigénétiques, qui peuvent cependant être transmises de manière héréditaire, ou de la transmission de molécules telles que des anticorps via le cytoplasme maternel, qui constitue aussi une forme d'hérédité non génétique (Danchin *et al.* 2011)

2. Notons que la plasticité phénotypique peut être adaptative, si les variations phénotypiques induites par l'environnement permettent à un génotype donné de maintenir la meilleure *fitness*.

douce (Albert *et al.* 2008 ; Chan *et al.* 2010), et d'autre part les allèles à dits à faibles effets, dont la présence sur un plus grand nombre de locus aboutit à l'augmentation de la valeur adaptative, comme c'est le cas pour certains traits continus comme la forme ou la taille chez divers animaux (voir notamment Savolainen *et al.* (2013) pour une revue).

## L'adaptation locale

Lorsque l'hétérogénéité de l'environnement induit une variation géographique des pressions de sélection, on fait alors référence à l'adaptation locale (Tiffin et Ross-Ibarra 2014). Une population localement adaptée aura alors la meilleure *fitness* dans cet environnement que n'importe quelle autre population introduite (Savolainen *et al.* 2013). Lorsqu'une population se voit confrontée à un nouvel environnement, un allèle favorable peut être sélectionné soit à partir d'une nouvelle mutation, dite *de novo*, soit à partir du polymorphisme déjà présent au sein de la population d'origine (*standing genetic variation*). Dans le premier cas, le temps de l'adaptation peut être plus long : en effet, il faudra d'une part attendre l'apparition de la mutation dans la population, et d'autre part sa fixation par sélection naturelle se fera à partir de fréquences plus faibles que dans le cas de la *standing genetic variation* (Barrett *et al.* 2008 ; Bock *et al.* 2015). Cependant, Colautti et Lau (2015) suggèrent que le grand nombre de traits polygéniques existants, s'ils sont associés à une grande taille de population, offrent un fort potentiel d'apparition *de novo* de mutations adaptatives.

Les allèles sélectionnés au cours de l'adaptation locale se distinguent en fonction de leur influence sur la valeur sélective dans les différents environnements (Tiffin et Ross-Ibarra 2014). Dans un cas, l'avantage que confère un allèle dans un environnement donné devient pénalisant dans un autre environnement, on parle alors de pléiotropie antagoniste[1] ou de compromis (*trade-off*) génétique. Dans l'autre cas, celui dit de la neutralité conditionnelle, un allèle possède un avantage seulement dans un environnement donné, et demeure neutre partout ailleurs (Figure 1). Tiffin et Ross-Ibarra (2014) suggèrent que la neutralité conditionnelle d'un allèle pourrait être plus fréquente qu'observée à l'heure actuelle, notamment car les test utilisés y sont moins sensibles (voir 1.2).

1. Ce terme doit sont origine à George C. Williams (1957) qui l'a utilisé pour la première fois afin de désigner les gènes impliqués dans le vieillissement qui sont à la fois bénéfiques avant la reproduction mais dont l'effet devient délétère par la suite

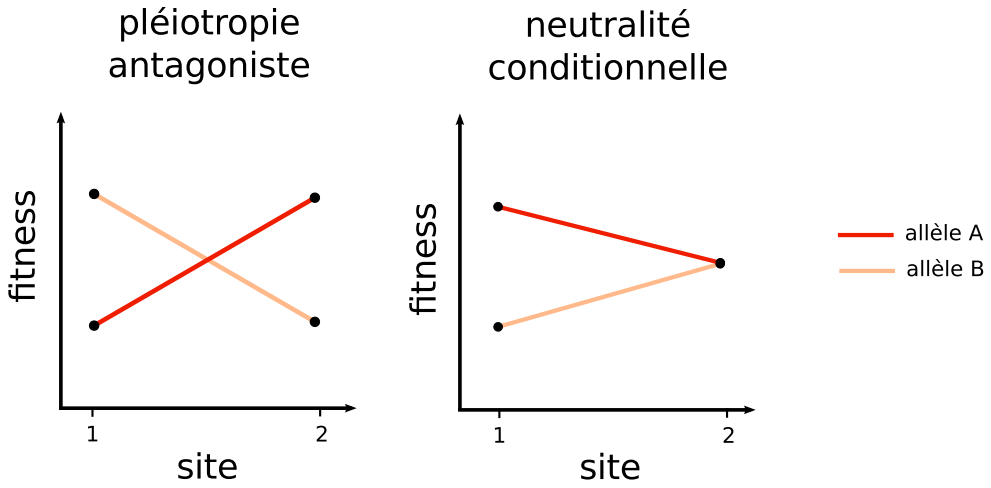



**Figure 1** – Effet de la pléiotropie antagoniste (à gauche) ou de la neutralité conditionnelle sur la *fitness* pour un locus à deux allèles. Dans le premier cas, chaque allèle permet de maximiser la *fitness* dans un environnement distinct (site 1 ou 2). Dans le cas de neutralité conditionnelle, l'un des deux allèles présente un avantage dans un environnent donné (l'allèle A au site 1), tandis qu'ailleurs les deux allèles sont équivalents (site 2).

Le maintien du polymorphisme génétique associé à l'adaptation locale est influencé par l'intensité du flux de gènes qui existe entre les populations qui subissent des pressions de sélection distinctes. En effet, si la migration d'individus entre les populations peut permettre l'introduction de variation génétique adaptative, elle peut aussi réduire la vitesse de fixation d'un allèle favorable (Sexton *et al.* 2014). Si le flux de gènes est intense, la sélection devra être suffisamment forte pour contrebalancer les effets de la migration. Ainsi, lorsque la migration est forte, il est attendu que l'adaptation locale soit majoritairement causée par un nombre restreint de locus à fort effet (Savolainen *et al.* 2013). Dans ce contexte, les allèles présentant un pléiotropisme antagoniste seront les plus à même de maintenir la différenciation génétique au(x) locus impliqué(s) dans l'adaptation locale. L'intensité de la sélection peut elle même être responsable de la restriction du flux de gènes, par exemple en contre-sélectionnant systématiquement tout migrant venant d'un environnement différent. L'isolement des populations contribuera alors à l'accentuation globale de la différenciation génétique.

Cependant, les facteurs régulant l'intensité du flux de gène peuvent être relativement indépendants du processus adaptatif, comme l'isolement par la distance lié aux capacités de dispersion de l'espèce, la présence de barrières géographiques, ou le contraste entre les environnements, qui peut être soit défavorable (en imposant par exemple un décalage des périodes de reproduction) soit favorable au flux de gène (vent, courants) (Sexton *et al.* 2014).

## 1.2 Détection des évènements d'adaptation sur les génomes

La sélection naturelle et les mutations qu'elle favorise ne sont bien entendu pas les seules forces évolutives à façonner l'architecture génomique : la migration est, comme nous l'avons vu précédemment, elle aussi importante, tout comme les variations aléatoires de fréquences alléliques dues à la dérive génétique. Cependant, et contrairement aux autres forces évolutives, la sélection naturelle n'a pour cible que certains locus, parfois même un seul polymorphisme nucléotidique (ou SNP, *Single Nucleotide Polymorphism*) alors que les évènements de mutation, migration et dérive ont une influence globale en affectant pour une population donnée tous les locus du génome avec la même probabilité (Lewontin et Krakauer 1973). C'est ce contraste entre évènements locaux et globaux à l'échelle du génome qui sont exploités par un grand nombre de méthodes visant à détecter des locus sous sélection.

Après avoir décrit l'empreinte caractéristique laissée par l'adaptation sur les génomes en parallèle des principales méthodes permettant de les détecter, nous nous attardons sur celles des scans génomiques basés sur la différenciation, qui permettent de rechercher des traces de sélection directionnelle chez des espèces non modèles.

### Le balayage sélectif et ses conséquences

Au cours d'un évènement d'adaptation, le succès reproducteur des variants génétiques dont les phénotypes sont en adéquation avec leur environnement est maximisé. Les mutants à l'origine de ce succès verront donc leur fréquence augmenter dans la population jusqu'à éventuellement atteindre la fixation. De part leur proximité physique (même chromosome), les allèles présents aux locus voisins du locus sélectionné vont eux aussi voir leur fréquence augmenter dans la population par le phénomène dit d'auto-stop génétique ou *genetic hitchhiking* (Maynard Smith et Haigh 1974). Il en résulte une réduction du polymorphisme génétique à proximité de la cible de la sélection ; la recombinaison génétique à l'œuvre chez les espèces sexuées, réduit quant à elle cet effet à mesure de l'éloignement avec l'allèle sélectionné (Figure 2). Ce phénomène est appelé balayage sélectif ou *selective sweep* (Nielsen 2005 ; Stephan 2015).

L'intensité et l'amplitude du balayage sélectif dépendent de la force de la sélection, qui résulte de la taille d'effet de l'allèle sélectionné (coefficient de sélection $s$), de la taille efficace [1] de la population ($N_e$), ainsi que du taux de recombinaison local. D'autre part, l'effet du *selective sweep* est maximum lorsqu'il est provoqué par la fixation rapide d'une mutation apparue *de novo*. En effet, la diversité génétique est rapidement réduite par l'invasion dans la population d'un haplotype (*i.e.* la combinaison linéaire des différents

1. La taille ou effectif efficace est une valeur théorique correspondant au nombre d'individus d'une population idéale (d'après le modèle de Wright-Fisher : taille constante, panmixie, générations non-chevauchantes) participant réellement à la reproduction ; elle permet notamment de mesurer l'influence de la dérive génétique qui sera d'autant plus importante que $N_e$ est petite

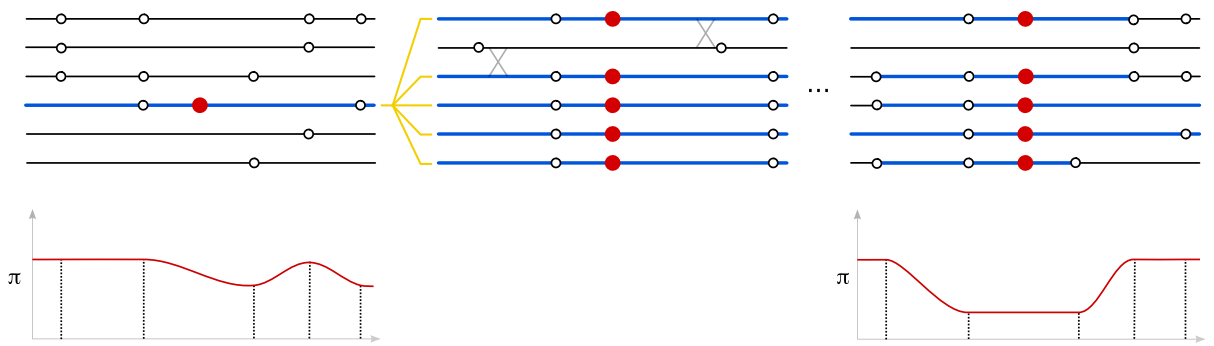



**Figure 2** – Effet d'un balayage sélectif sur la diversité nucléotidique $\pi$. Lorsqu'un allèle favorable apparaît dans la population (points rouges en haut à gauche) les polymorphismes neutres (dont les points blancs représentent un allèle dérivé [1]) à proximité sont emportés par auto-stop génétique ce qui provoque une baisse de la diversité génétique. La recombinaison (croix grises) limite quant à elle l'ampleur de cette baisse au voisinage du site sélectionné (en bas à droite), laissant apparaître une "signature" de la sélection sur le génome.

allèles portés par un chromosome) où la recombinaison et l'apparition de mutations n'ont pas encore disposé d'assez de temps pour rétablir le polymorphisme. On parle alors de *hard sweep*. Dans le cas où la mutation sélectionnée est recueillie à partir de la *standing genetic variation*, un plus grand polymorphisme est attendu dans les régions voisines de l'allèle favorable, car celui-ci existe depuis plus longtemps dans la population, et son haplotype original a pu donc subir plus d'évènements de recombinaison et acquérir de nouvelles mutations. Sa sélection se fait donc à partir d'un plus grand nombre d'haplotypes, il s'agit alors d'un *soft sweep*. Quand certains gènes ou régions du génome sont suspectés d'avoir évolué de manière adaptative, il est possible de comparer la diversité génétique de ces candidats à celle d'autres locus supposés neutres. Sous l'hypothèse d'un balayage sélectif, la diversité génétique du locus candidat devrait être significativement réduite. Ceci étant, une simple variation des taux de mutations le long du génome peut aussi bien expliquer ces différences. Ces variations étant cependant bien conservées entre espèces, on peut alors distinguer ces évènements en comparant les ratios entre substitutions nucléotidiques intra-spécifique (polymorphisme) et inter-spécifique (divergence) à différents locus. Ce test, qui porte le nom de HKA en référence à ses auteurs Hudson, Kreitman et Aguadé (1987), suppose que sous un régime neutre, le ratio polymorphisme sur divergence est constant entre les différents locus ; une déviation significative de ce ratio à un locus, du fait de la baisse du polymorphisme, sera alors le signe d'un *selective sweep*.

La réduction de la diversité génétique à proximité de la cible de la sélection au cours d'un balayage sélectif conduit par ailleurs à l'augmentation locale des valeurs de déséquilibre de liaison. Des tests basés sur la mesure au sein d'une population d'un haplotype anormalement long à haute fréquence (*e.g. Extended Haplotype Homozygosity* [EHH], Sabeti *et al.* (2002) ou *integrated Haplotype Score* [iHS], Voight *et al.* (2006)), permettent

1. Un allèle dérivé est le produit d'une mutation par rapport à l'état allélique ancestral, qui peut être inféré en faisant appel à un groupe externe (Fay et Wu 2000)

alors d'inférer l'action d'un balayage sélectif (Oleksyk *et al.* 2010).

Une autre signature du balayage sélectif est le changement de la distribution (ou spectre) des fréquences des allèles dérivés (*site frequency spectrum*, SFS). Celui-ci donne, pour une population donnée la proportion des mutations retrouvées $i = 1, 2, ..., n-1$ fois au sein de l'échantillon $n$. Sous un modèle d'évolution neutre et à l'équilibre, cette proportion est décroissante et proportionnelle à $1/i$ (Nielsen 2005). On trouve donc plus souvent des allèles dérivés partagés par un faible nombre d'individus. Dans le cas d'un *selective sweep*, l'élimination d'une partie de la diversité génétique tend à biaiser le SFS vers une proportion plus importante de mutations à faibles et fortes fréquences, il "creuse" en quelque sorte le spectre de fréquence des allèles dérivés (Figure 3). En effet, ce sont alors les quelques haplotypes n'ayant pas été éliminés ainsi que l'action, même faible de la recombinaison qui conduira à observer une fréquence proche de $1/n$ ou $(n-1)/n$ aux sites neutres par rapport à l'état ancestral (Fay et Wu 2000).

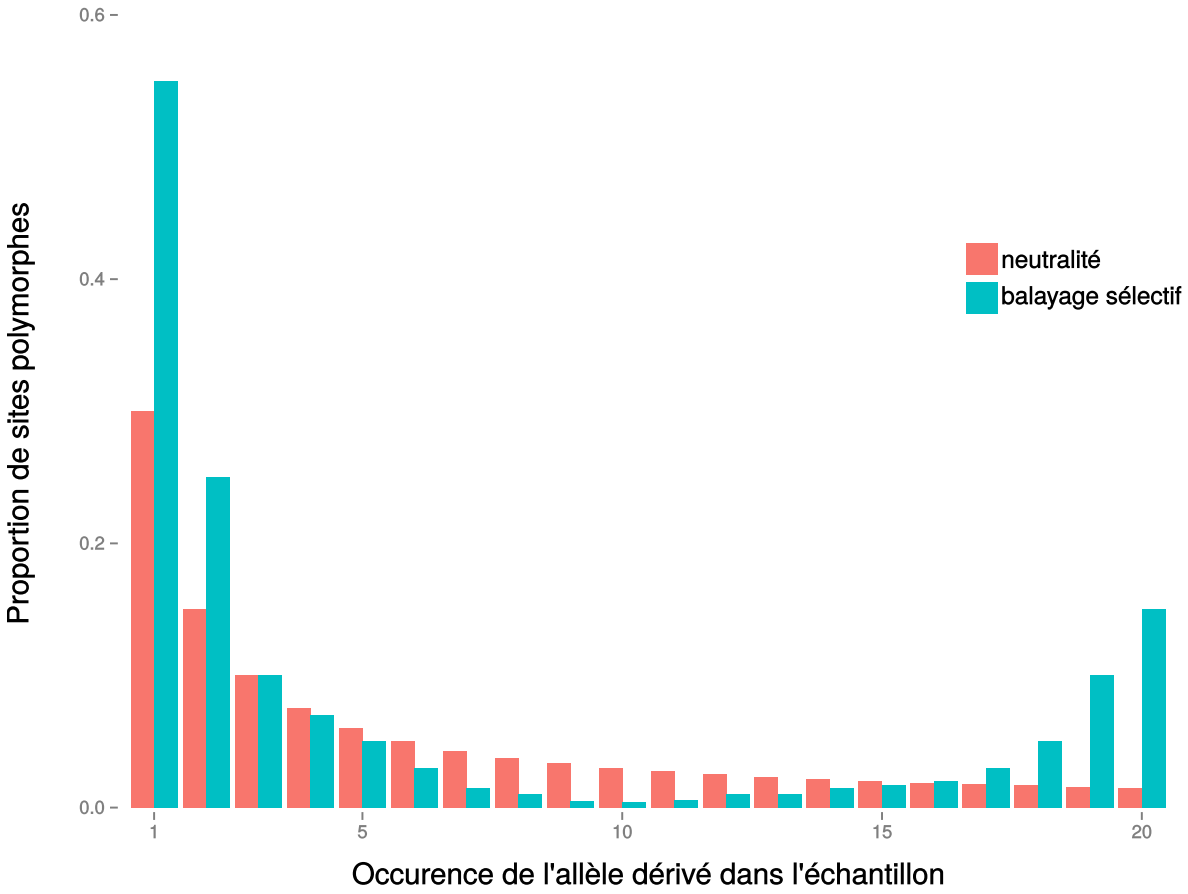



**Figure 3** – Spectre des fréquences des allèles dérivés dans les cas de neutralité sélective ou de balayage sélectif. D'après (Nielsen 2005)

De nombreux test de sélection sont basés sur la modification du SFS par les *selective sweep*. Le plus connu est certainement le test du $D$ de Tajima (1989), qui permet notamment de révéler l'excès de mutations à faibles fréquences. Cependant, une telle observation

est aussi compatible avec une récente diminution de la taille de la population [1] (*bottleneck*) ; il peut ainsi être intéressant de combiner les résultats du $D$ de Tajima aux valeurs du test $H$ de Fay et Wu (2000), et qui est lui sensible à l'augmentation de la proportion d'allèles dérivés à fortes fréquences (Holsinger 2012).

En réduisant localement la diversité génétique, il est aussi attendu qu'un balayage sélectif conduise à une augmentation de la différenciation génétique (variation des fréquences alléliques) entre des populations adaptées à des environnements différents à proximité du locus sélectionné (Stephan 2015). Dans le cadre de cette thèse, nous allons nous intéresser en détail aux méthodes basées sur cette signature, qui s'affranchissent de certaines contraintes imposées par celles décrites précédemment. Elles permettent notamment, *via* l'utilisation de marqueurs génétiques répartis le long du génome, de rechercher sans *a priori* des traces d'adaptation locale. Il n'est donc pas nécessaire de disposer de locus candidats ou de faire appel à un groupe externe (*e.g.* test HKA, iHS), ou encore de disposer de séquences homologues complètes ($D$ de Tajima, $H$ de Fay et Wu) ou même d'un génome de référence (EHH, iHS).

### Les scans génomiques basés sur la différenciation

La différenciation représente la distance séparant des populations sur la base de leur composition génétique. Elle peut être notamment mesurée en quantifiant leurs variations en termes de fréquences alléliques. Dans ce cas, la différenciation est souvent calculée à partir de l'indice de fixation de Wright, le $F_{ST}$. Celui-ci correspond à la réduction d'hétérozygotie par rapport l'attendu, du fait de la structuration de la population totale en sous-populations (ST pour Subpopulation/Total). Les valeurs du $F_{ST}$ s'échelonnent théoriquement entre 0 – les fréquences alléliques sont les mêmes entre sous-populations considérées – et 1 – chaque sous-population a fixé un allèle différent –. Au niveau des polymorphismes neutres, la différenciation génétique mesurée entre populations est principalement due à deux facteurs. Le premier correspond à l'échantillonnage statistique, du fait que seulement une partie des individus de chacune des populations est prélevée. Le second correspond à l'équilibre entre dérive génétique (échantillonnage génétique) et migration ; la première ayant tendance à accroître le $F_{ST}$ alors que la migration tend à homogénéiser les fréquences alléliques entre les populations. Un balayage sélectif cependant, peu localement augmenter le $F_{ST}$ au delà des valeurs observées par la simple balance dérive/migration ou les effets d'échantillonnage. Les locus présentant une telle différenciation sont qualifiés d'*outliers* et ont alors une forte probabilité d'être soit la cible de la sélection, soit d'avoir été entrainés par autostop-génétique (Figure 4).

1. cet effet devrait alors être global, et donc rendre plus difficile la détection d'un *selective sweep*

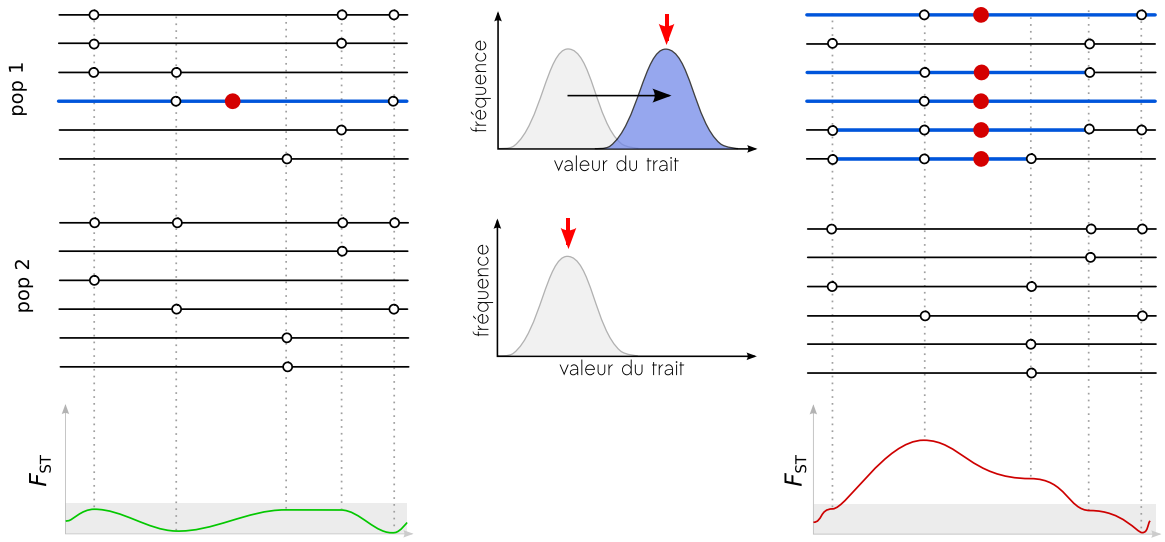



**Figure 4** – Principe du scan génomique basé sur la différenciation génétique. A gauche, avant la sélection de l'allèle favorable dans la population 1 (pop1) la différenciation génétique ($F_{ST}$) mesurée entre les deux populations correspond à la balance dérive/migration et l'échantillonnage. La variation des pressions de sélection en faveur d'un nouveau phénotype (flèches rouges au centre), conduit à l'augmentation en fréquence de l'haplotype (représenté par une ligne bleue) associé à l'allèle favorable dans la population 1. A la suite du balayage sélectif, la réduction de diversité génétique des locus à proximité de la cible de la sélection contribue à augmenter la différence de fréquences alléliques entre les populations et conduit à l'augmentation de la différenciation génétique, au-delà des valeurs moyennes (zone grisée).

Sur ce principe, Lewontin et Krakauer (1973) ont introduit le premier test permettant de rejeter la neutralité à un locus donné si la valeur de différenciation mesurée dépasse un seuil critique estimé par leur modèle. Cependant, le seuil utilisé par Lewontin et Krakauer a rapidement été discuté, certains aspects démographiques comme la migration, les variations de taille efficace ou la structuration des populations pouvant engendrer des valeurs bien supérieures au sein de la distribution neutre à celles estimées par les auteurs (Robertson 1975 ; Nei et Maruyama 1975).

L'incorporation de modèles démographiques plus complexes aux méthodes de scans génomiques (Figure 5) a permis le développement de méthodes basées sur la simulation d'une "enveloppe neutre" de différenciation à partir des locus échantillonnés. Par exemple, le modèle `Fdist` (Beaumont et Nichols 1996) simule la différenciation neutre entre différentes populations (dèmes) qui échangent entre elles des migrants à un taux déterminé (modèle en îles de Wright ; voir Figure 5 A). Les valeurs de $F_{ST}$ observées au delà d'un seuil critique relatif à l'hétérozygotie des locus (par exemple supérieures au 95ème centile) peuvent être attribuées à l'action de la sélection directionnelle. Si cette méthode s'avère robuste à certains écarts par rapport au modèle démographique initial (par exemple la colonisation en pas japonais ou *stepping-stone*), les résultats peuvent s'avérer lourdement biaisés si les taux de migrations entre dèmes diffèrent (Beaumont et Nichols 1996 ; Vitalis *et al.* 2001 ; Foll et Gaggiotti 2008). Afin d'éviter cet écueil, Vitalis *et al.* (2001) ont proposé d'utiliser un modèle plus simple (Figure 5 B) simulant l'évolution par dérive seule des locus de deux populations n'échangeant pas de migrants, mais issues d'une population

ancestrale pouvant avoir subi un évènement de *bottleneck* (package `DETSEL`). En réalisant l'analyse par paires de populations, on s'affranchit des problèmes de covariance des fréquences alléliques entre certaines populations si la structure ne reflète pas strictement un modèle en îles. En limitant l'analyse à deux échantillons, la méthode pèche en revanche par un manque de puissance statistique ; et la multiplication des comparaisons par paires augmente la probabilité de faux positifs, car chaque paire considérée n'est pas réellement indépendante des autres (Beaumont 2005).

Une alternative aux modèles de scans génomiques basés sur la simulation d'une enveloppe neutre est l'estimation directe, via une méthode bayésienne[1], de la probabilité qu'un locus a d'être sous sélection (ou d'être entrainé par balayage sélectif). La méthode ainsi utilisée par Foll et Gaggiotti (2008) repose sur un modèle général de fission où les différents dèmes dérivent indépendamment d'une population ancestrale, d'où ils peuvent avoir reçu des proportions inégales de migrants (Figure 5 C). Ce modèle, qui peut être apparenté à celui en îles de la méthode `Fdist`, présente cependant l'avantage d'avoir des hypothèses relâchées tolérant différents taux de migration et différents effectifs de populations. Suivant l'approche proposée par Beaumont et Balding (2004), le test de Foll et Gaggiotti (2008) (implémenté dans le package `Bayescan`) décompose la différenciation ($F_{ST}$) entre un dème et la population ancestrale en deux paramètres : le premier, $\beta$ , incorpore les effets neutres (dérive, migration) et est commun à tous les locus de la population. Le second, $\alpha$ , inclut, lui, l'effet de la sélection qui est propre à chaque locus ($\alpha$ est alors un paramètre partagé pour un locus entre toutes les populations). Pour chaque locus, `Bayescan` estime la probabilité postérieure $P(\alpha \neq 0 \mid \text{données})$ permettant de le désigner comme *outlier*. Une limite commune à ces modèles, est l'absence de prise en compte d'une possible structure hiérarchique des populations. Ce type de structure peut s'observer si les flux de gènes sont restreints entre certains groupes de populations, ou si les populations étudiées sont issues de pools génétiques distincts. Elle peut notamment provoquer une corrélation des fréquences alléliques neutres entre certains dèmes, conduisant à l'augmentation de la variance du $F_{ST}$ entre dèmes issus de pools de migrants différents. De tels écarts aux modèles initiaux peuvent être alors à l'origine d'une quantité importante de faux positifs (Lotterhos et Whitlock 2014). Afin de prendre en compte cette structure hiérarchique dans le flux de gènes, Excoffier *et al.* (2009) ont proposé une extension du modèle de Beaumont et Nichols (1996) en effectuant la simulation de l'enveloppe neutre d'après un modèle en îles hiérarchiques (logiciel `Arlequin3.5` ; Figure 5 D). Plus récemment, les méthodes bayésiennes ont elles aussi incorporé ce modèle dans de nouvelles implémentations (Foll *et al.* 2014 ; Duforet-Frebourg *et al.* 2014).

Un aspect important de ces méthodes concerne la nature des marqueurs employés.

1. Une méthode bayésienne repose sur l'estimation d'une distribution de paramètres (dite postérieure) à partir d'une distribution supposée (prior) en maximisant leur vraisemblance avec les données observées

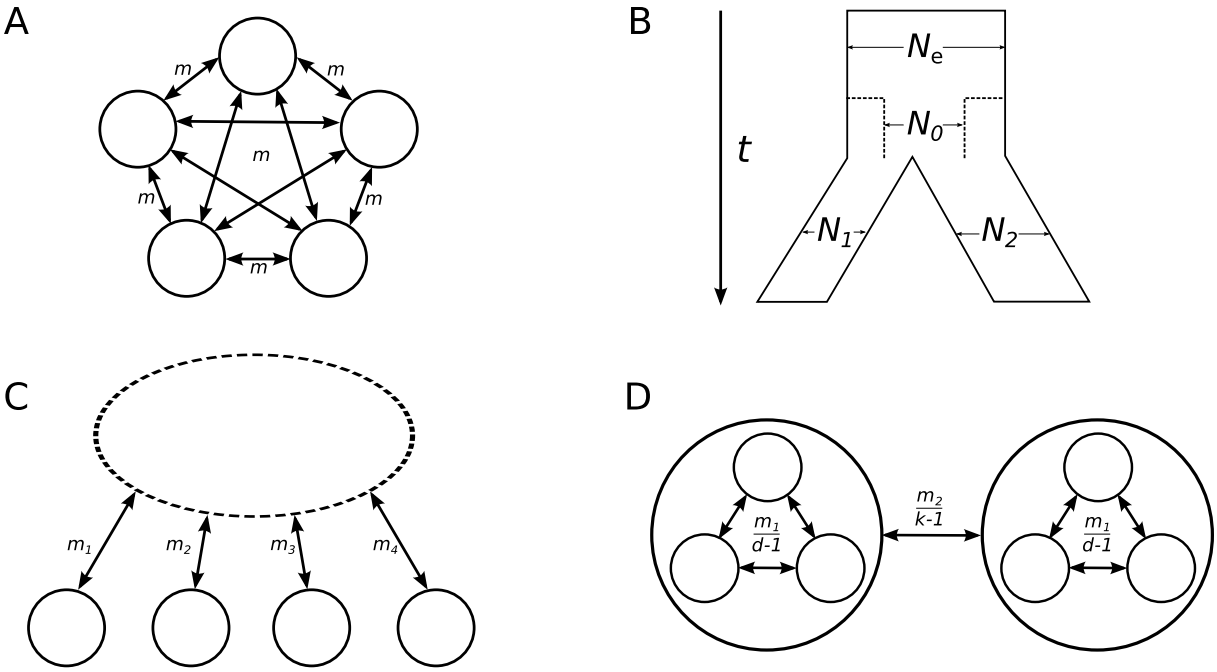



**Figure 5** – Modèles démographiques utilisés par les principales méthodes de scans génomiques basés sur la différenciation. A : `Fdist/dFdist` modèles en îles de Wright ; des populations de taille identique échangent entre elles des migrants au même taux $m$. B : `DETSEL/DETSELD` Deux populations de taille constante $N_1$ et $N_2$ divergent par pure dérive (pas de migration) après leur séparation depuis une population ancestrale de taille $N_0$, qui peut avoir subi un évènement de *bottleneck* après une période à l'équilibre mutation/dérive avec une taille constante $N_e$. Le temps $t$ est représenté par la flèche verticale de gauche. C : `Bayescan` modèle de fission ; les différentes populations divergent d'une population ancestrale dont elles ont pu recevoir une quantité différente de migrants (taux $m_1$, $m_2$, $m_3$, $m_4$). Ce modèle peut être considéré comme équivalent au modèle en îles de Wright avec des taux de migration et des tailles de populations variables. D : `Arlequin 3.5` modèle en îles hiérarchiques ; au sein de chaque groupe les sous-populations échangent des migrants avec le même taux $m_1/(d-1)$ avec $d$ le nombre de sous-populations ; des migrants peuvent aussi être échangés entre groupes au taux $m_2/(k-1)$ avec $k$ le nombre de groupes.

Le choix des marqueurs résulte souvent d'un équilibre qu'il convient de trouver entre leur niveau d'information, leur densité à travers le génome, leurs taux et types de mutation, la facilité technique et financière de les génotyper et enfin leur prise en charge par les méthodes statistiques existantes. En effet, les modèles présentés précédemment ont principalement été conçus pour un usage avec des marqueurs utilisés au moment de leur développement (Figure 6), à savoir les microsatellites, les AFLP (*Amplified Framgment-Length Polymorphism*) et les SNPs.

Les microsatellites sont des marqueurs très utilisés en génétique des populations. Ils sont co-dominants, et leurs différents allèles correspondent au nombre de répétitions d'un motif di- tri- ou tetra- nucléotidique, qu'il est possible d'amplifier sur chacun des chromosomes à l'aide d'une paire d'amorces spécifiques situées de part et d'autre des répétitions (Ellegren 2004). Leur taux de mutation (gain ou perte de répétitions) élevé fait que ces marqueurs sont pertinents pour relater les évènements les plus récents de l'évolution génétique des populations. Leur amplification et génotypage est relativement aisé, et du fait de leur fort polymorphisme (certains locus peuvent avoir des dizaines d'allèles) une

simple dizaine de locus microsatellites peut permettre de décrire avec précision la structure des populations. En revanche le développement d'un nombre suffisamment élevé pour une étude de scan génomique est fastidieux et devient rapidement coûteux, en particulier pour des espèces non modèles. En effet une paire d'amorce doit être utilisée par locus, et il est nécessaire de disposer de ressources génomiques (génome de référence, séquençage) afin de les identifier. Malgré tout, la recherche de traces de d'adaptation à l'aide de ces marqueurs a pu être réalisée encore récemment (Gu *et al.* 2009 ; Meier *et al.* 2011 ; Jaquiéry *et al.* 2012).

Le scan génomique basé sur la différentiation s'est aussi popularisé grâce aux méthodes permettant de produire rapidement et sans connaissances particulières sur le génome des centaines de marqueurs. Au cours des années 2000, les AFLP ont alors suscité un fort engouement. Cette méthode utilise des enzymes de restriction afin de découper le génome ; des adaptateurs sont ligués aux extrémités des fragments de restriction et une PCR, réalisée à l'aide d'une seule paire d'amorces, permet d'amplifier jusqu'à une centaine de bandes qu'il est possible de discriminer par électrophorèse (Vos *et al.* 1995). La combinaison de plusieurs jeux d'enzymes et d'amorces permet alors de générer des centaines de marqueurs pour le scan génomique. L'inconvénient principal de cette méthode est qu'il s'agit de marqueurs dominants, c'est à dire qu'il est impossible de distinguer les génotypes homo- et hétérozygotes pour la présence du site de restriction responsable de l'amplification de la bande. L'accès direct aux fréquences alléliques, utilisées pour calculer le $F_{ST}$, est donc impossible. Il reste possible d'inférer ces fréquences, par exemple en faisant l'hypothèse que les populations sont à l'équilibre de Hardy-Weinberg, ou en disposant d'une estimation du coefficient de consanguinité pour chacune des populations ($F_{IS}$) précédemment obtenue à l'aide d'un marqueur co-dominant (Zhivotovsky (1999), voir Bonin *et al.* (2007) pour une liste exhaustive des autres méthodes). Dans leur version pour marqueurs dominants, `Fdist` (`DFdist`, publiée plus tard dans le package `MCHEZA`, Antao et Beaumont (2011)) et `DETSEL` (`DETSELD`, non publiée) implémentent la méthode bayésienne de Zhivotovsky (1999). `Bayescan` peut aussi être utilisé avec des marqueurs dominants ; dans ce cas, il intègre l'incertitude quant aux fréquences alléliques en laissant le $F_{IS}$ (réduction d'hétérozygotie due à la consanguinité) varier librement entre 0 et 1 au cours de l'estimation des paramètres (Foll et Gaggiotti 2008). Une comparaison de ces trois méthodes utilisant des AFLP (ou autres marqueurs dominants) préconise l'utilisation de `Bayescan`, qui détecte le plus de locus réellement sous-sélection tout en réduisant le taux de faux positifs (Pérez-Figueroa *et al.* 2010). En définitive, l'avantage certain des AFLP en termes de coût et d'investissement par rapport aux microsatellites, leur a assuré un succès qui n'a été compromis que par l'arrivée de nouvelles méthodes permettant de génotyper directement des milliers de substitutions nucléotidiques (SNP).

Si les SNP, en général bi-alléliques, sont *per se* moins informatifs que d'autres marqueurs co-dominants comme les microsatellites, ils constituent les polymorphismes les plus

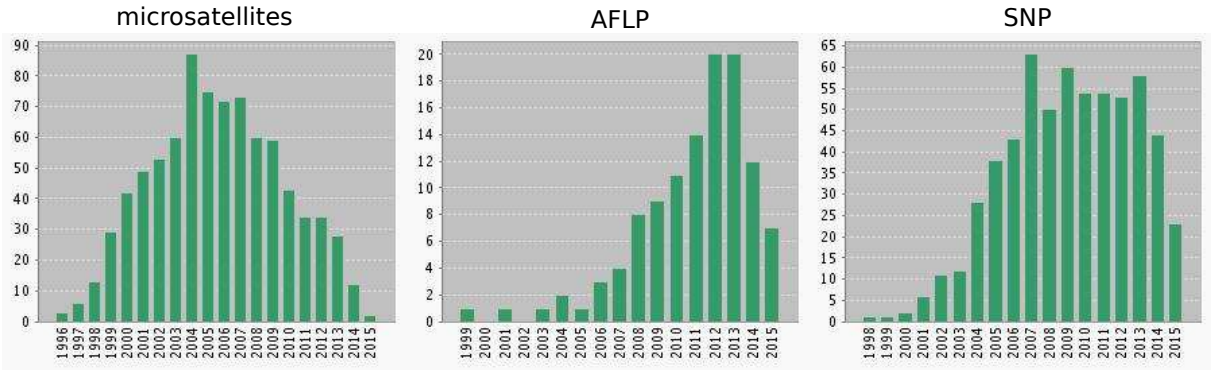



**Figure 6** – Nombre annuel de publications référencées sur la plateforme Web of Science (Thomson Reuters) incluant les termes *à gauche* "microsatellite" ; *au centre* ("AFLP" OR "Amplified fragment length polymorphism") ; *à droite* ("SNP" OR "RAD*seq*") et ("genom* scan" OR "selection scan" OR "selection fingerprint" OR "selection * signature" OR "signature * selection"). Ces chiffres illustrent la transition progressive opérée entre l'utilisation des microsatellites puis des AFLP, remplacés aujourd'hui par les SNP.

denses au sein des génomes. Leur génotypage a récemment été facilité chez les espèces non modèles par l'avènement de la méthode de RAD-seq (*Restriction site Associated DNA sequencing*, Miller *et al.* 2007). En combinant la digestion enzymatique de l'ADN total avec le séquençage à haut débit des régions directement adjacentes au site de restriction (les "*RAD-tags*" ou locus RAD) il est possible de générer plusieurs milliers de marqueurs SNPs sans connaissance *a priori* du génome. Le RAD-seq, ainsi que d'autres méthodes basées sur l'identification de SNPs (*e.g.* capture de séquences, re-séquençage) bénéficient du coût réduit du séquençage à haut débit et font que ces marqueurs sont certainement aujourd'hui les plus populaires dans les expériences de scans génomiques.

### Limites des approches de scan génomique

Si les effets démographiques sont de mieux en mieux pris en compte dans les méthodes actuelles de scan génomique, d'autres facteurs confondants et certaines limitations sont à prendre en compte lors de l'interprétation des résultats.

Les *soft sweeps* qui ont lieu notamment sur des locus à faible effet, par exemple lors de sélection polygénique, ou encore si le variant sélectionné provient de la *standing genetic variation* ainsi que les balayages partiels, se produisant si l'allèle favorable n'a pas atteint la fixation, sont plus difficilement détectables par les méthodes de scan génomique. L'effet du balayage sélectif étant réduit, il est plus difficile de le mettre en évidence par des méthodes "*outlier*" classiques. De la même manière, les allèles adaptatifs qui présentent une neutralité conditionnelle (voir 1.1) produiront des valeurs de différenciation moins importantes que dans le cas des allèles ayant une pléiotropie antagoniste (Tiffin et Ross-Ibarra 2014). Une approche prometteuse est celle employée par Daub *et al.* (2013, 2015) : celle-ci ne cherche plus à déterminer si un locus en particulier est *outlier*, mais si un groupe de locus candidats associés au sein d'une même voie métabolique présentent un niveau de

différenciation significativement différent du reste du génome. En revanche, cette méthode requiert de disposer de solides ressources génomiques sur le modèle étudié (ici l'Homme) et reste donc limitée à quelques espèces.

Les processus sélectifs à l'œuvre au sein des génomes peuvent eux mêmes perturber la détection d'événements de sélection positive. Le premier exemple est celui de la sélection d'arrière plan (*Background Selection* BGS) : celle-ci correspond au phénomène d'élimination des mutations délétères (sélection négative, ou purifiante) et a comme conséquence de réduire la diversité génétique aux sites neutres qui y sont liés (Charlesworth *et al.* 1993). Cet effet peut alors être confondant avec celui d'un balayage sélectif, et le sera d'autant plus dans les régions de faible recombinaison (Stephan 2010). La BGS a par ailleurs des effets confondants avec les processus démographiques sur le SFS (notamment une expansion de population) ; sa prise en compte conjointe à ces effets peut alors être nécessaire afin de réduire les taux de faux positifs (Bank *et al.* 2014).

Le second exemple concerne la fréquence des balayages sélectifs. Il a notamment été montré que lorsque ceux-ci se produisent de manière récurrente, cela peut avoir pour effet de réduire la proportion d'allèles dérivés à fortes fréquences, dont l'excès est une signature typique des *selective sweep* sur le SFS (Kim 2006). Par ailleurs, lorsque des balayages sélectifs se produisent de manière simultanée, l'interaction entre ces évènements peut avoir comme conséquence de restaurer une partie de la diversité génétique, encore une fois d'autant plus si un tel phénomène a lieu dans des régions de faible recombinaison (Stephan 2015).

Le taux de recombinaison est également un facteur important, notamment car celui-ci n'est pas toujours homogène au sein des génomes : il peut être fortement réduit à proximité des centromères, ou sur les hétérochromosomes et parfois fortement accru dans des régions appelées *hotspots*, positionnées le long des chromosomes (Nachman 2002 ; McVean *et al.* 2004 ; Kulathinal *et al.* 2008). Ainsi, et particulièrement lorsque des marqueurs "anonymes", c'est à dire dont la localisation n'est pas connue a priori, sont utilisés, il est possible que des locus*outliers* se trouvent en réalité très éloignés des cibles réelles de la sélection en raison d'un faible taux de recombinaison. A contrario, l'empreinte du balayage dans les *hotspots* sera limité et ceux-ci seront donc plus difficiles à identifier.

Une autre source de confusion lors de scans génomiques, en particulier basés sur la différenciation, est la superposition des barrières dites endogènes au flux de gène (c'est à dire indépendantes de l'environnement comme des incompatibilités génétiques) et les barrières dites exogènes, comme celles induites par la sélection divergente liée à l'environnement (Bierne *et al.* 2011). Ces barrières génétiques étant souvent semi-perméables, elles augmentent la variance des valeurs de différenciations mesurées le long du génome. Par exemple, l'hybridation inter-spécifique peut entraîner une forte différenciation génétique aux locus introgressés entre les populations n'ayant pas vécu les mêmes évènements d'hybridation. Dans le cas d'un scan génomique, si des zones de contact inter-spécifiques

se superposent aux barrières écologiques (induisant l'adaptation) les effets seront confondants (Fraïsse *et al.* 2015). Une solution est donc, si possible, de disposer de réplicats géographiques des conditions de l'environnement pour lesquels l'adaptation est testée (Fraïsse *et al.* 2015).

Les différents écueils évoqués ici doivent rappeler que ces tests ne servent qu'à relever des indices, compatibles avec les traces laissées par la sélection naturelle sur les génomes. L'étape suivante et nécessaire est l'identification de locus candidats à l'adaptation (gènes, séquences régulatrices), soit dans l'environnement génétique lié soit éventuellement directement au site *outlier*. Il pourra être alors nécessaire de mettre en évidence la réduction caractéristique du polymorphisme génétique associée au balayage sélectif, par exemple en séquençant de manière intensive la région génomique associée à un *outlier*. Là encore, l'identification, par exemple de gènes dont la fonction est compatible avec un scénario adaptatif ne constitue pas une preuve, sinon une piste, qu'il conviendra d'éprouver afin de relier enfin une variation génétique à un phénotype adaptatif.

Ainsi, des approches de génomique fonctionnelle pourront être envisagées comme la mise sous silence ciblée de gènes candidats, par exemple via l'utilisation des méthodes d'ARN interférence (dégradation ciblée des transcrits par l'utilisation de petits ARN spécifiques, Kim et Rossi (2008)). Une autre méthode, très prometteuse, est celle d'édition du génome par l'utilisation de la protéine bactérienne CRISPR-Cas9 permettant de générer des mutations ciblées (clivage double brin) à l'aide d'un ARN guide (Ran *et al.* 2013). Suite à ces modifications, des mesures de *fitness* en conditions contrôlées (jardin commun) permettront enfin de relier la variation génétique au phénotype adaptatif.

## 1.3 Génétique de l'invasion et adaptation locale

Une espèce est aujourd'hui considérée comme invasive si celle-ci est introduite dans un nouvel environnement et engendre des dégâts de nature économique, environnemental, ou encore si elle représente une menace pour l'homme (Beck *et al.* 2008). Si les invasions biologiques ne sont pas récentes, l'intensification des échanges mondiaux liés à l'activité humaine a fortement contribué à leur multiplication (Keller *et al.* 2014). Cependant, et en dépit de leur impact considérable sur les écosystèmes et les sociétés, les espèces invasives représentent une opportunité unique d'étudier *in situ* l'adaptation locale. En effet, les espèces que nous considérons aujourd'hui comme invasives sont celles qui ont traversé avec succès les différents filtres menant à leur implantation pérenne dans un nouvel environnement, et l'on peut alors s'intéresser au rôle du processus adaptatif dans ce succès. Il est important de noter que même si l'environnement colonisé apparait très semblable à celui d'origine, un certain nombre de nouvelles interactions biotiques et abiotiques est susceptible d'induire de nouvelles pressions de sélection, favorisant l'adaptation locale (Colautti et Lau 2015).

L'invasion de nouveaux territoires par une espèce modifie son patrimoine génétique d'une part en raison des effets démographiques – dont l'influence est globale sur le génome – et d'autre part à cause de l'apparition de nouvelles pressions de sélection – qui ne concernent elles probablement qu'un nombre réduit de locus–. Il est donc important d'avoir à l'esprit certains attendus, à la fois théoriques et empiriques concernant ces processus particuliers, rassemblés sous le terme de "génétique de l'invasion".

### Effets fondateurs, invasions multiples et diversité génétique

Le principal attendu dans le cas des invasions biologiques, est l'observation d'un effet fondateur qui peut être considéré comme un cas particulier de goulot d'étranglement, ou *bottleneck* (réduction drastique de la taille efficace de la population) pouvant se produire lorsqu'une partie de la population ancestrale est prélevée puis introduite dans un nouvel environnement. Cela peut avoir pour conséquence de réduire la diversité génétique de la population invasive et donc la probabilité d'y sélectionner un variant favorable. Par ailleurs, si la taille efficace reste réduite à l'issue d'un *bottleneck*, l'influence de la dérive génétique rendra difficile toute emprise de la sélection naturelle, et donc tout processus adaptatif.

A cela peut s'ajouter un effet lié à l'action de la dérive au cours de l'expansion géographique : l'*allele surfing* (Klopfstein *et al.* 2006 ; Excoffier et Ray 2008). En effet, comparée à son centre, la densité de la population se trouve fortement réduite aux limites de l'aire de répartition, ce qui peut entrainer la fixation aléatoire et locale d'allèles qui vont alors "surfer" sur le front de migration (Figure 7). Les conséquences peuvent être la fixation de mutation délétères, on parle alors de charge d'expansion (*expension load*) (Peischl *et al.*

2013 ; Peischl et Excoffier 2015). D'autre part, elle peut conduire à la détection de fortes valeurs de différenciation mesurées entre des échantillons géographiquement distincts à certains locus. Ces valeurs de différentiation pourraient alors être confondues avec des traces d'adaptation locales lors des scan génomiques (Hofer *et al.* 2009).

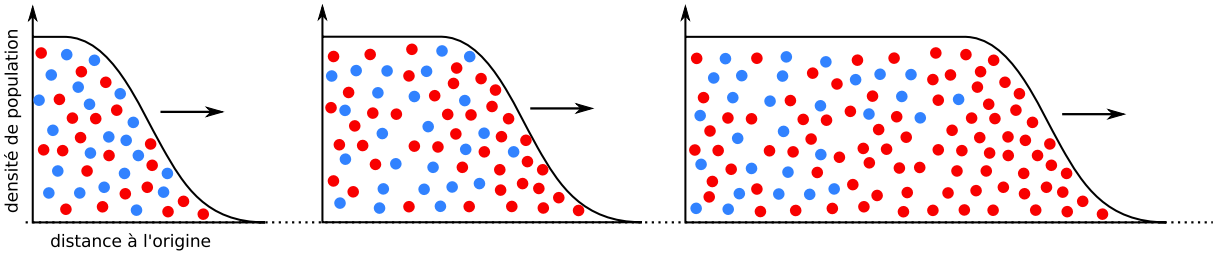



**Figure 7** – Fixation aléatoire d'une mutation par *allele surfing* dans le cas d'un locus à deux allèles (représentés par les points bleus et rouges). La densité de population réduite en marge de l'aire de répartition augmente localement a dérive génétique et peut conduire à la fixation d'un allèle (ici le rouge), qui "surfe" sur le font de migration.

Cependant, les nombreux évènements d'invasion biologique ont pu montrer que les introductions multiples (*i.e.* répétées et/ou impliquant des sources différentes) pouvaient être très fréquentes, ce qui peut aboutir à l'observation d'une diversité génétique égale, voire supérieure de la population invasive par rapport à celle observée dans l'aire d'origine (Handley *et al.* 2011). En effet, la pression de propagules, c'est à dire le nombre absolu d'individus qui arrivent dans le nouvel environnement, correspond directement à l'effort d'échantillonnage génétique réalisé dans l'aire d'origine (Bock *et al.* 2015). Enfin, les *bottlenecks* évacuent principalement les allèles rares et peuvent dans certains cas préserver une partie de la diversité génétique ; de fait, ils ne seront alors pas forcément un frein à l'adaptation locale des espèces invasives (Dlugosch et Parker 2008 ; Dlugosch *et al.* 2015).

### Les moyens de l'adaptation locale

Dans une récente revue sur le rôle de la diversité génétique dans le succès des espèces invasives, Dlugosch *et al.* (2015) soutiennent que leur réussite dépend plus largement du type de diversité génétique introduite que de sa quantité.

Par exemple, les locus à forts effets (décris en 1.1), qui ont par définition une forte réponse par rapport à la sélection, ont été souvent découverts chez des espèces invasives (Dlugosch *et al.* 2015) ; en effet cette propriété augmente la probabilité de fixation de l'allèle favorable lorsque la taille efficace de la population est fortement réduite. Par ailleurs, dans ces situations où l'espérance de vie d'un allèle à faible fréquence est fortement réduite, des allèles favorables provenant de la variation génétique préexistante atteindront plus rapidement la fixation qu'une mutation *de novo* (Bock *et al.* 2015). Parmi eux, ceux présentant une neutralité conditionnelle ont une plus forte probabilité de ségréger à des fréquences intermédiaires dans les populations natives et donc de mieux résister aux effets aléatoires causés par la démographie lors de l'invasion (Dlugosch *et al.* 2015).

Les introductions multiples sont aussi l'opportunité de créer, via le contact de populations initialement allopatriques, de nouvelles combinaisons génétiques favorables dans le nouvel environnement (Handley *et al.* 2011 ; Bock *et al.* 2015 ; Colautti et Lau 2015 ; Dlugosch *et al.* 2015). L'expression de vigueur hybride ou la purge des mutations délétères parfois associés aux évènements d'hybridation intra-spécifiques sont autant d'avantages pour le maintien d'une espèce invasive dans sa nouvelle aire de répartition (Bock *et al.* 2015).

Il a également été montré que les inversions chromosomiques avaient pu permettre à certaines espèces d'augmenter de manière drastique leur aire de répartition *via* la fixation de combinaisons génétiques favorables (Dlugosch *et al.* 2015 ; Kirkpatrick et Barrett 2015). Ces mutations, parfois acquises lors d'hybridations inter-specifiques, font partie des locus à large effet décrits plus haut, également nommés "super-locus". De tels réarrangements chromosomiques sont notamment associés à la colonisation de nouveaux environnements chez le moustique *Anopheles gambiae* (adaptation aux climats arides en Afrique via l'hybridation avec *Anopheles arabiensis*, Fouet *et al.* (2012)), et les Drosophiles *D. melanogaster* (Australie, Hoffmann et Weeks (2007)) et *D. subobscura* (Amériques, Prevosti *et al.* (1988) ; Ayala *et al.* (1989)). En plus de l'acquisition directe d'un génotype adaptatif, les inversions permettent notamment de prévenir la destruction de la combinaison allélique favorable par la recombinaison génétique (Kirkpatrick et Barrett 2015). Elles peuvent représenter cependant un coût à la sélection, notamment chez les hétérozygotes dont la fertilité est parfois réduite. La modélisation a récemment montré que l'impact des inversions chromosomiques sur la fertilité, même faible, peut annihiler son potentiel adaptatif (Dlugosch *et al.* 2015).

La réduction voir l'absence de variabilité génétique au moment de l'invasion, pourtant nécessaire à l'adaptation locale, peut en certains cas être compensée par des mécanismes permettant "d'attendre" l'arrivée de variation génétique. La plasticité phénotypique peut par exemple être un trait transitoirement sélectionné lors de l'invasion ; les différents phénotypes produits par la plasticité peuvent par la suite être associés à de nouvelles variations génétiques, au cours d'un processus appelé assimilation génétique (Lande 2015). Le domaine de l'épigénétique permet de comprendre les modifications, parfois héritables, issues de l'interaction des gènes et de leur produits (méthylation, marques d'histones). Des travaux récents suggèrent que de telles variations induites par l'environnement peuvent faciliter l'adéquation d'une espèce avec celui-ci en l'absence de variabilité nucléotidique. Ce phénomène pourrait tout de même s'apparenter à une forme d'adaptation (Bock *et al.* 2015).

# 2 Éléments Transposables et génomique des populations

## 2.1 Rapide tour d'horizon des Éléments Transposables (ET)

Les ET sont des séquences génomiques mobiles, capables de s'insérer à de nouveaux locus au cours d'un processus appelé transposition.

Leur présence au sein des génomes est quasi-universelle et inclut les procaryotes (on parle alors d'IS pour *Insertion Sequences*) et les eucaryotes animaux et végétaux. Leur nature répétée contribue largement à la taille des génomes, qui est très bien corrélée à la proportion d'ET chez les eucaryotes (Lynch et Conery 2003 ; Biémont et Vieira 2004 ; Chénais *et al.* 2012).

Les ET peuvent être classés en fonction de leur mode de transposition. Ainsi, les ET de Classe I ou rétrotransposons transposent selon un mode de "copier-coller" ; ils utilisent un intermédiaire à ARN qui sera rétro-transcrit avant d'être intégré à un nouveau locus. Les éléments de Classe II, appelés parfois transposons ou éléments à ADN, se déplacent principalement par "couper-coller". Cette opération nécessite une excision puis une réintégration à l'aide d'une transposase (Wicker *et al.* 2007).

L'ensemble des copies d'une même séquence constitue une *famille*, dont la séquence type, ou consensus, peut faire partie d'une *super-famille*, inclue dans un *clade* appartenant à l'une ou l'autre deux classes évoquées précédemment.

Les ET regroupent en réalité une grande diversité de séquences, de quelques dizaines à plusieurs milliers de paires de bases, et incluent souvent un ou plusieurs cadres de lecture ouverts leur permettant d'encoder les enzymes nécessaires à leur transposition. Certains ET, comme les SINE (Short INterspersed Elements [Classe I]) ou les MITE (Miniature Interspersed Transposable Element [Classe II]) en sont dépourvus, et sont alors dépendant des machineries encodées par d'autres ET ; ils sont alors qualifiés d'éléments non autonomes.

De par leur dynamique au sein des génomes, les ET sont à l'origine de différents types de mutations. Ils peuvent s'insérer dans des gènes ou séquences régulatrices, en perturber l'expression (agir comme activateurs ou inhibiteurs lorsqu'ils sont situés à proximité), ou encore du fait de leur répétition le long des chromosomes, agir comme points d'appuis de recombinaisons éctopiques (recombinaisons entre régions chromosomiques non homologues), provoquant d'importants remaniements chromosomiques (Casacuberta et González 2013).

Les ET représentent ainsi une part non négligeable de la variabilité génétique produite au cours de l'évolution. Par exemple, chez*Drosophila melanogaster*, il a pu être montré que 80% des mutations à effet phénotypique étaient dues à l'activité des éléments transposables (Green 1988). Par ailleurs, leur importance est d'autant plus remarquée lorsque certaines

familles d'ET sont spécifiquement mobilisées en présence d'un stress environnemental (*e.g.* choc thermique) ou génomique (comme l'hybridation inter-spécifique) (Capy *et al.* 2000 ; Biémont et Vieira 2006). Bien qu'à l'image de toute mutation l'activité des ET peut être souvent neutre, ces évènements de transposition peuvent être tout aussi bien délétères (Goodier et Kazazian 2008 ; Beck *et al.* 2011 ; Vela *et al.* 2014) qu'à l'origine d'un certain nombre de phénotypes adaptatifs (Casacuberta et González 2013 ; Stapley *et al.* 2015).

## 2.2 Les ET marqueurs génétiques

L'activité des ET façonne le polymorphisme génétique entre espèces, populations et individus. La transmission verticale (de parents à descendants) des insertions acquises au cours de l'évolution peut ainsi être utilisée pour mesurer le degré d'apparentement entre individus : la présence d'un même ET au même locus chez deux individus représente en effet une très forte probabilité d'identité par descendance. Le polymorphisme d'insertion de différentes familles d'ET peut ainsi être utilisé comme marqueur génétique, au même titre que les microsatellites, AFLP ou autre SNP.

Différentes méthodes ont été développées afin de mettre en évidence ces variations. Si une grande majorité des exemples concerne des espèces végétales (chez qui les ET peuvent être très abondants), l'utilisation de ces méthodes chez les animaux, et en particulier chez les moustiques, a permis de traiter différentes questions de génétique des populations comme l'étude de la structure (Barnes *et al.* 2005 ; Boulesteix *et al.* 2007 ; Esnault *et al.* 2008) ou la recherche de trace de sélection par scan génomique (Bonin *et al.* 2008).

Le génotypage du polymorphisme d'insertion repose sur la capture de séquences dont la présence et/ou la taille sont caractéristiques d'une insertion. Le point commun des différentes méthodes qui y sont dédiées consiste à mettre en évidence les polymorphismes sans *a priori* sur leur localisation au sein du génome. Pour des familles d'ET très répétées, il est par exemple possible d'amplifier par PCR les fragments se situant entre deux ET adjacents à l'aide d'amorces complémentaires de leur séquence. La taille finale de l'amplicon est spécifique de l'insertion de deux copies de la famille d'ET utilisée) (méthode *Inter-Retrotransposon Amplification Polymorphism* – IRAP ou *Inter-MITE Polymorphism* – IMP Grzebelus 2006).

Il est aussi possible de capturer systématiquement chaque insertion par Southern blot. Après digestion enzymatique de l'ADN, les fragments sont séparés en fonction de leur taille par électrophorèse, puis ceux associés aux insertions d'ET sont sélectionnés par hybridation de sondes spécifiques aux familles d'ET étudiées (*e.g.*, Boulesteix *et al.* 2007).

Alternativement au Southern blot, et à la manière d'une AFLP, des adaptateurs peuvent être ligués aux sites de coupures. Un couple d'amorces ciblant une portion de l'ET et la séquence de l'adaptateur permet d'amplifier par PCR un grand nombre des insertions d'une famille considérée (méthode *Sequence-Specific Amplification polymorphism*

S-SAP, Grzebelus 2006). Cette procédure peut être rendue plus spécifique (notamment afin d'éviter les amplifications adaptateur-adaptateur) en enrichissant les produits de PCR avec les séquences contenant l'ET. Pour cela, il est possible de réaliser un marquage de l'amorce ET à la biotine puis d'effectuer une capture par streptavidine des fragments amplifiés (Witherspoon *et al.* 2010) ; une autre méthode consiste à réaliser deux PCR successives avec des amorces nichées dans la séquence de l'élément (Esnault *et al.* 2008). Enfin, cet enrichissement peut aussi être achevé en utilisant des adaptateurs incomplets ou en "Y" ; ces constructions sont réalisées à partir de deux oligomères dont une partie seulement (celle permettant la liaison au site de restriction) sont complémentaires ; à l'autre extrémité, l'un des brins comporte une séquence identique à l'amorce destinées à l'adaptateur et devra donc être complémentée lors de la première élongation de la PCR, de l'ET vers l'adaptateur, avant que son amorce ne s'y hybride (Carnelossi *et al.* 2014). Ces méthodes dérivées du S-SAP sont communément regroupées sous le terme de Transposon Diplay (TD).

Deux limites peuvent néanmoins être associées à ces méthodes. Premièrement, ces marqueurs sont de types dominant dans la plupart des cas. . Les états homozygotes ou hétérozygotes d'une insertion sont en effet indiscernables, ce qui pose en génétique des populations le problème de l'inférence des fréquences alléliques. Heureusement, l'essor des AFLP a largement contribué au développement d'un cadre statistique permettant de prendre en compte de tels marqueurs et notamment au cours de scan génomiques (voir 5). Cependant, la seconde critique concerne le fait que les méthodes de biologie moléculaire décrites précédemment ne permettent pas en l'état de générer les centaines, voire les milliers, de marqueurs nécessaires à un scan génomique.

Ainsi quelques études se sont consacrées à la tâche difficile d'optimiser ces méthodes vers des approches à plus haut débit. Bonin *et al.* (2008) ont par exemple combiné une approche de TD avec une méthode de génotypage sur puce à ADN DArT (pour *Diversity Array Technology*) : au cours d'une première étape, les insertions amplifiées par TD sont clonées pour un groupe d'individus représentatifs. Ces clones DArT sont ensuite fixés sur une puce. Le protocole de TD est ensuite répété pour l'ensemble des individus et les produits de PCR sont hybridés sur la puce à ADN, permettant de révéler la présence ou l'absence de chacune des insertions de la librairie. Cette approche a permis de générer 500 marqueurs polymorphes utilisés par la suite lors d'un scan génomique chez le moustique *Aedes aegypti* (Bonin *et al.* 2009). Cette méthode limite cependant le nombre de polymorphismes étudiés dans la population totale à ceux détectés au sein de l'échantillon utilisé pour la construction de la librairie ; de plus la construction de celle-ci implique un effort supplémentaire de clonage préalable pour chacune des insertions étudiées.

Afin de s'affranchir de ces contraintes, le séquençage direct sur plateforme à haut débit des produits de TD peut être une solution. En combinant un TD biotine/streptavidine au séquençage en Illumina (Witherspoon *et al.* 2010) ou un TD avec adaptateurs "incomplets"

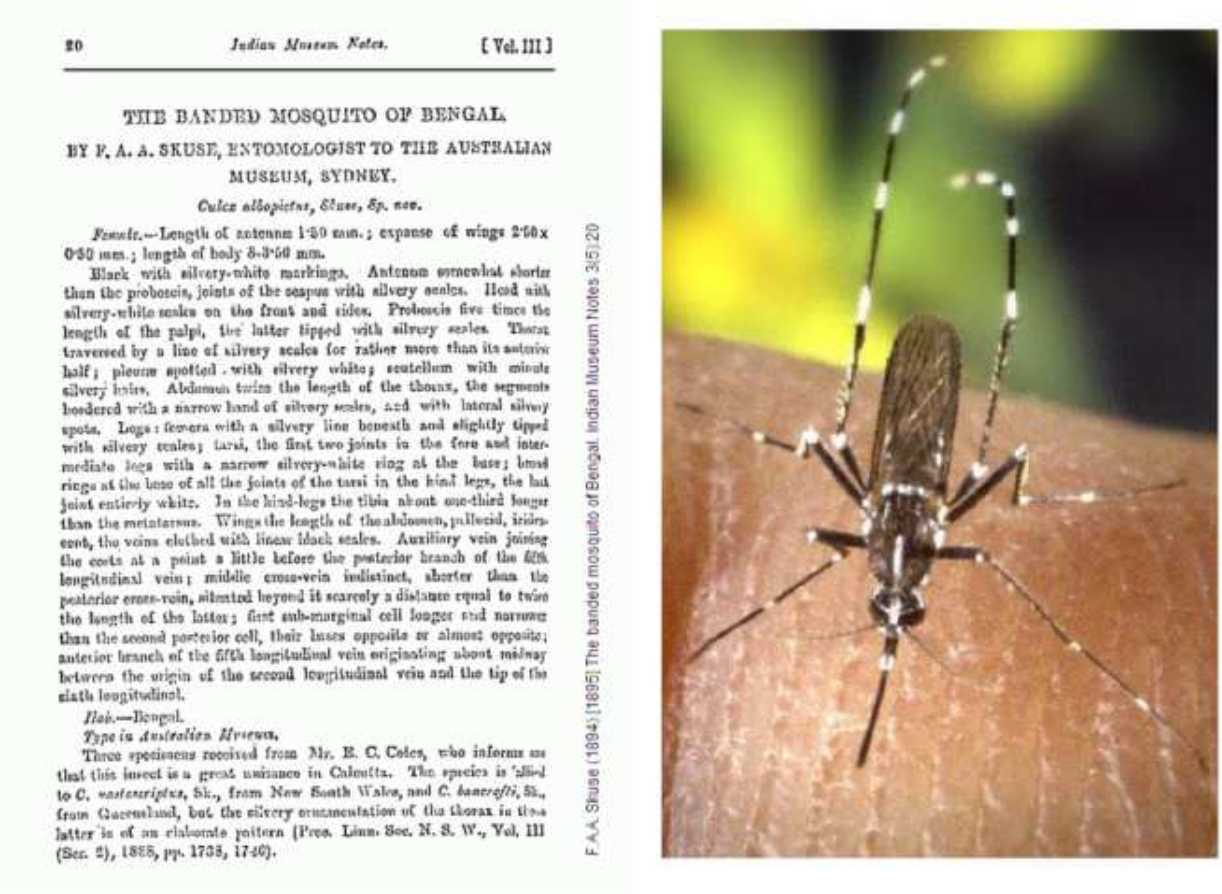
au séquençage sur 454 (Iskow *et al.* 2010) il a été rendu possible d'étudier le polymorphisme d'insertion des familles très répétées *Alu* et *L1* chez l'Homme (plusieurs milliers de copies).

Cette approche est particulièrement séduisante, notamment avec l'utilisation de la technologie de séquençage par paires de lectures (Illumina *Paired-end*), qui permet d'obtenir pour chacune des insertions amplifiées une paire de séquences comportant d'une part un fragment de l'ET et de l'autre une portion de locus flanquant. En l'absence de génome de référence, à la manière des approches développées en RAD-sequencing pour reconstituer les locus RAD, il est possible de reconstituer les locus d'insertion par clustering des reads appartenant à la région flanquante, et ensuite d'identifier pour chacun des individus la présence ou non de chacune des insertions, sans connaissance génomique préalable. En choisissant des familles d'ET très répétées et relativement bien conservées pour que la PCR du TD soit possible à l'aide d'une seule paire d'amorce par famille, il est donc envisageable de générer facilement une collection importante de marqueurs.

# 3 Le moustique tigre *Aedes albopictus*

## 3.1 Biologie descriptive

Le moustique tigre *Aedes* (Stegomya) *albopictus* (Skuse 1894), ou *Asian tiger mosquito* en anglais, est un Diptère (insecte) de la famille des Culicidae, sous-famille Culicinae appartenant à la tribu des Aedini et aux sous-genre *Stegomya*. Il se distingue ainsi d'autres Culicinae tels que les moustiques du genre *Culex*, et de par son appartenance à une autre sous-famille, des espèces du genre *Anopheles* (anophèles). Il est décrit pour la première fois par Skuse au Bengale (Inde) sous le nom de *Culex albopictus*, "le moustique bagué du Bengale".



20 *Indian Museum Notes.* [Vol. III]

THE BANDED MOSQUITO OF BENGAL.

BY F. A. A. SKUSE, ENTOMOLOGIST TO THE AUSTRALIAN MUSEUM, SYDNEY.

*Culex albopictus, Skuse, Sp. nov.*

*Female.*—Length of antennæ 1·50 mm.; expanse of wings 2·50 x 0·80 mm.; length of body 3·3·50 mm.

Black with silvery-white markings. Antennæ somewhat shorter than the proboscis, joints of the scapus with silvery scales. Head with silvery-white scales on the front and sides. Proboscis five times the length of the palpi, the latter tipped with silvery scales. Thorax traversed by a line of silvery scales for rather more than its anterior half; pleuræ spotted with silvery white; scutellum with minute silvery hairs. Abdomen twice the length of the thorax, the segments bordered with a narrow band of silvery scales, and with lateral silvery spots. Legs: femora with a silvery line beneath and slightly tipped with silvery scales; tarsi, the first two joints in the fore and intermediate legs with a narrow silvery-white ring at the base; broad rings at the base of all the joints of the tarsi in the hind legs, the last joint entirely white. In the hind-legs the tibia about one-third longer than the metatarsus. Wings the length of the abdomen, pellucid, iridescent, the veins clothed with linear black scales. Auxiliary vein joining the costa at a point a little before the posterior branch of the fifth longitudinal vein; middle cross-vein indistinct, shorter than the posterior cross-vein, situated beyond it scarcely a distance equal to twice the length of the latter; first sub-marginal cell longer and narrower than the second posterior cell, their bases opposite or almost opposite; anterior branch of the fifth longitudinal vein originating about midway between the origin of the second longitudinal vein and the tip of the sixth longitudinal.

*Hab.*—Bengal.

*Type in Australian Museum.*

Three specimens received from Mr. E. C. Coles, who informs us that this insect is a great nuisance in Calcutta. The species is allied to *C. vastaseriptus*, Sk., from New South Wales, and *C. bancrofti*, Sk., from Queensland, but the silvery ornamentation of the thorax in the latter is of an elaborate pattern (Proc. Linn. Soc. N. S. W., Vol. III (Ser. 2), 1888, pp. 1735, 1746).

F.A.A. Skuse (1894) [1895] The banded mosquito of Bengal. Indian Museum Notes 3(5):20

**Figure 8** – Extrait de l'article original de Skuse (1894) [à g.] décrivant "*Culex albopictus*" (domaine public) ; [à d.] femelle adulte lors d'un repas de sang sur un hôte humain (CDC/James Gathany, domaine public)

### Cycle de vie

Le moustique tigre est un insecte anautogène, c'est à dire que la femelle réalise un repas de sang afin d'assurer la production de ses œufs. Le délai entre le repas de sang et l'oviposition est compris entre 2 et 4 jours (Hawley 1988). La femelle pond ses œufs juste au-dessus de la surface de l'eau ; au laboratoire, ceux-ci éclosent entre 2 et 5 jours

après immersion pour des températures comprises entre 24 ° C et 30 ° C (voir aussi Figure 9), cette durée pouvant atteindre 10 jours à plus faible température (Hawley 1988). Les œufs d'*Ae. albopictus* sont capables de résister à la dessiccation par une phase dite de quiescence, avant le retour des conditions environnementales nécessaires à leur éclosion ; celle-ci pouvant atteindre 2 à 4 mois (Hawley 1988 ; Poelchau *et al.* 2013c).

Les populations présentes dans les environnements tempérés peuvent également initier une phase de diapause : en réponse aux changements de la photopériode et de la température à l'approche de l'hiver, la femelle pond un œuf dans lequel la larve entrera en dormance (Mori *et al.* 1981). La diapause, à la différence de la quiescence, ne sera levée qu'après une période déterminée, et ce malgré le retour des conditions favorables (Mori *et al.* 1981 ; Hawley 1988 ; Poelchau *et al.* 2013c), il s'agit en quelque sorte d'une quiescence "programmée". Ces aspects sont notamment importants dans le statut invasif d'*Ae. albopictus* (voir 1.2.2).

Après l'éclosion, le cycle larvaire voit se succéder 4 stades ponctués de mues et aboutit à la métamorphose lors d'un ultime stade de nymphe (pupe). Les stades larvaires et la métamorphose s'étendent respectivement sur des périodes observées entre 5 et 10 jours et entre 1 et 3 jours, ces durées étant inversement proportionnelles à des températures entre 20 et 30 degrés (Hawley 1988). Les mâles présentent par ailleurs un temps de développement larvaire inférieur à celui des femelles, et ce même dans des conditions réduites de nourriture (Hawley 1988). En conséquence, l'émergence des mâles se fait avant celle des femelles. Bien que la reproduction n'intervienne pas avant les premières 24h post-émergence de ces dernières, les accouplements ont principalement lieu dans les 5 à 10 premiers jours de l'imago, en vol et souvent à proximité des hôtes utilisés pour le repas de sang (Ali et Rozeboom 1973 ; Estrada-Franco et Graig 1995 ; Boyer *et al.* 2011). Le sperme du mâle est stocké dans les spermathèques, assurant à la femelle une fécondation des œufs tout au long de sa vie. Si les femelles peuvent s'accoupler à plusieurs reprises, les secrétions provenant des glandes accessoires des mâles, transférées avec le sperme, leur assurent la paternité de la descendance lors des cycles gonotrophiques suivants si tant est que la femelle n'est pas fécondée par un nouveau mâle dans un intervalle de 40 minutes (Oliva *et al.* 2013). Le cycle gonotrophique, qui correspond au temps entre deux ovipositions, dure environ 5 jours, mais là encore sa durée peut varier en fonction de la température ou de la lignée étudiée (Hawley 1988).

Durant le stade imago, le moustique se nourrit de liquides sucrés, principalement de nectar de fleurs (Hawley 1988 ; Müller *et al.* 2011). La longévité des femelles est supérieure à celle des mâles. Au laboratoire, celle-ci est mesurée entre 6 à 8 semaines contre 2 à 3 pour les mâles (Hawley 1988). Dans la nature, cette mesure obtenue à partir de diverses expériences de capture-marquage-recapture et soutenue par la modélisation serait d'environ une dizaine de jours (Hawley 1988 ; Brady *et al.* 2013).

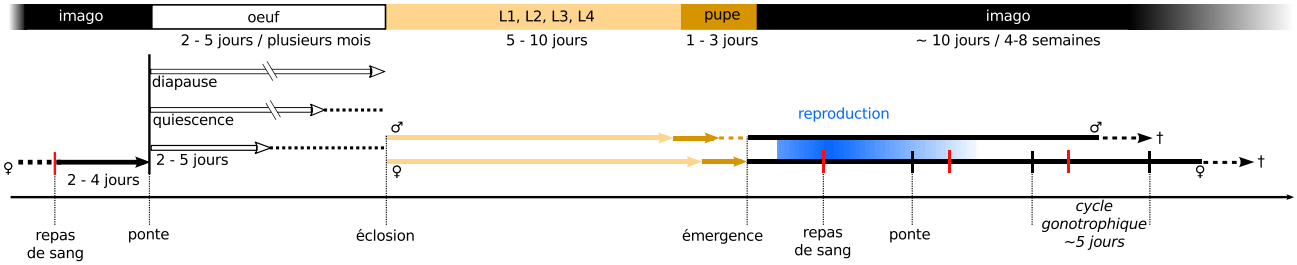



**Figure 9** – Cycle de vie du moustique tigre *Ae. albopictus*(voir texte)

## Ecologie

***Habitat.*** Il est communément admis que le moustique tigre est originaire des forêts du sud-est asiatique, en raison du fait que les larves des espèces appartenant à son sous-groupe (Albopictus) n'ont seulement été identifiées que dans des trous d'arbres de cette région (Hawley 1988). Au sein de son aire d'origine, l'espèce est cependant présente dans de nombreux environnements à la fois tropicaux (climat chaud et humide avec de faibles variations de températures annuelles) et tempérés (saisons contrastées incluant des températures négatives l'hiver, des pics de chaleur l'été et une plus faible humidité). *Ae. albopictus* est aujourd'hui souvent retrouvé dans les zones péri-urbaines ou rurales (Hawley 1988 ; Paupy *et al.* 2009). Sa présence dans les milieux anthropisés s'explique par l'utilisation d'une grande variété de sites de ponte à la fois naturels (trous d'arbres, bambous, feuilles des Broméliacées) et surtout artificiels (pneus, caisses, vases, coupelles,...). Ce type de sites de ponte fait que le moustique tigre est souvent retrouvé dans des zones d'activités industrielles (ports, aires de stockage, casses automobiles,...) ou moins actives comme les cimetières (Paupy *et al.* 2009 ; Bonizzoni *et al.* 2013). Cependant, contrairement à d'autres espèces antropohiles comme *Aedes aegypti*, le moustique tigre est principalement retrouvé à l'extérieur des habitations. La femelle favorise des sites de pontes sombres (potentiellement riches en matière organique) et proches du sol (Hawley 1988 ; Williges *et al.* 2014).

***Hôtes.*** Lorsqu'il est présent, l'Homme est aujourd'hui le principal hôte sur lequel la femelle *Ae. albopictus* réalise son repas de sang, ce qui a notamment été démontré par des expériences de choix ou l'analyse du contenu des repas de sang de populations naturelles (Paupy *et al.* 2009 ; Bonizzoni *et al.* 2013). L'Homme peut cependant être facilement remplacé dans les zones les moins anthropisées (Sivan *et al.* 2015) et ces analyses montrent par ailleurs que le spectre d'hôtes, qui s'étend principalement aux mammifères, peut aussi concerner des oiseaux, des amphibiens et des reptiles (Hawley 1988 ; Paupy *et al.* 2009 ; Bonizzoni *et al.* 2013).

***Dispersion naturelle.*** Plusieurs études basées sur des expériences de marquage-recapture sur le terrain ont permis d'estimer que les capacités naturelles de dispersion d'*Ae. albopictus* sont de l'ordre de quelques centaines de mètres (Hawley 1988 ; Niebylski et Craig 1994 ; Bellini *et al.* 2010 ; Liew et Curtis 2004 ; Marini *et al.* 2010). Le vol, pratiqué

proche du sol chez cette espèce, pourrait notamment limiter ses capacités de dispersion par le vent (Hawley 1988). Ainsi, si les distances maximum mesurées atteignent les 600 à 800 mètres, la plupart des données disponibles rapportent des moyennes inférieures à 400 mètres.

***Compétition***. Chez les moustiques la compétition a principalement lieu au stade larvaire. Celui-ci, inféodé au site de ponte, sera donc dépendant de la qualité du micro-habitat dans lequel l'œuf a été pondu. A l'échelle intra-spécifique, il a été montré que la présence d'œufs ou de larves sur un site de ponte était à l'origine d'un compromis évolutif : à faible densité, leur présence, qui peut être interprétée comme un gage de qualité du milieu, favorise la ponte, puis l'effet s'inverse avec l'augmentation de la densité initiale (Wasserberg *et al.* 2014). A l'échelle inter-spécifique, la compétition a été étudiée entre *Ae. albopictus* et différentes espèces de son aire de répartition actuelle ou envisagée dans le cadre de son invasion. De nombreuses études ont clairement montré un avantage de la larve d'*Ae. albopictus* par rapport à *A. aegypti*, la compétition pouvant avoir des conséquences asymétriques (c'est à dire plus sévères pour *Ae. aegypti*) sur la survie de l'adulte et aboutir à l'exclusion de cette espèce de certaines zones péri-urbaines ou rurales (Paupy *et al.* 2009 ; Bonizzoni *et al.* 2013), et(Bonizzoni *et al.* 2013 ; Alto *et al.* 2015). D'autres études ont par ailleurs indiqué un avantage des larves d'*Ae. albopictus* contre *Aedes sierrensis* (présent dans l'ouest des USA) (Kesavaraju *et al.* 2014), *Culex quinquefasciatus* (cosmopolite) (Allgood et Yee 2014) ou encore *Culex coronator* (Yee et Skiff 2014). Cependant, la majorité de ces travaux concerne des expériences en laboratoire ; la complexité de l'environnement naturel permet parfois d'observer la co-existance d'*Ae. albopictus* avec d'autres espèces – y compris *A. aegypti* – *via* l'utilisation de différents micro-habitats, ce qui peut alors faire varier l'intensité et les conséquences de la compétition (Bonizzoni *et al.* 2013).

***Microbiome***. La communauté microbienne du moustique tigre est largement dominée par la bactérie endosymbiotique *Wolbachia pipientis* ; qui atteint une prévalence proche de 100% dans les populations naturelles (Bourtzis *et al.* 2014) et peut représenter jusqu'à 99% des bactéries présentes chez les femelles (Minard *et al.* 2014). Deux souches : *w*AlbA et *w*AlbB ont été identifiées et sont la plupart du temps présentes en co-infection, qualifiée de "super-infection". Ces souches sont à l'origine de différents cas d'incompatibilités cytoplasmiques (IC), et l'importance de ce symbiote sera notamment discutée dans un contexte de génétique des populations dans l'article de revue (chapitre 2).

***Vecteur de pathogènes***. Comme de nombreux insectes hématophages, *Ae. albopictus* est impliqué dans la transmission de pathogènes. En laboratoire, le moustique tigre a été reconnu capable de transmettre 26 arbovirus (*arthropod born viruses*) parmi lesquels 14 ont été détectés sur des individus sauvages (Paupy *et al.* 2009). Ces virus incluent notamment 4 sérotypes (sur 5) de la dengue (DENV, flavivirus), le *West Nile virus* (WNV, flavivirus), le virus de l'encéphalite japonaise (flavivirus), le virus du Chi-

kungunya (CHIKV, alphavirus), le virus de l'encéphalite équine de l'Est (alphavirus) ou encore le virus de La Crosse (bunyavirus). Cependant son rôle effectif en tant que vecteur n'a seulement été formellement identifié que pour le DENV et le CHIKV, des virus transmissibles à l'Homme auxquels il doit en partie son statut d'espèce invasive (voir 3.2). *Ae. albopictus* est aussi impliqué dans la transmission naturelle de nématodes filaires du genre *Dilophilaria*, qui revêtent un intérêt vétérinaire (Paupy *et al.* 2009).

### Génétique

Trait extrêmement bien conservé chez les Culicidae[1], le caryotype du moustique tigre comprend trois paires de chromosomes (Hawley 1988). La paire 1 héberge le locus *Sex* (Mutebi *et al.* 1997), qui assure le déterminisme sexuel chez les Culicinae (Rai et Black 1999). Sa taille est inférieure aux chromosomes 2 et 3, de tailles identiques, un trait partagé au sein du genre *Aedes*. Les 3 paires sont métacentriques (Hawley 1988).

La taille du génome d'*Ae. albopictus* pourrait être extrêmement variable en fonction des populations : des mesures par cytophotométrie de Feulgen réalisées par Rao et Rai (1987) et Kumar et Rai (1990) affichent des valeurs comprises entre 0,62 et 1,66 pg pour une cinquantaine de lignées à travers le monde. Bien que la méthode employée à l'époque suscite des critiques (Greilhuber 1998, 2005), des estimations basées sur la cinétique de ré-association de l'ADN ont confirmé le patron de variation (une différence de x2) pour deux populations précédemment étudiées (Black et Rai 1988). Par ailleurs, cette étude montre que ces variations sont principalement dues à des différences dans la quantité d'ADN répété, dont l'hétérogénéité des proportions avaient déjà été observée entre plusieurs lignées d'*Ae. albopictus* (McLain *et al.* 1987). Au cours de cette thèse, nous avons mesuré la taille du génome d'une souche provenant de la Réunion par cytométrie de flux et obtenu une valeur de 1,19 pg soit environ 1,16 Gpb (voir chapitre 2).

1. A l'exception de *Chagasia bathana* qui possède 3 paires d'autosomes et une paire de chromosomes sexuels hétéromorphes (Rai et Black 1999)

## 3.2 *Ae. albopictus* : une espèce invasive

Le moustique tigre est aujourd'hui considéré comme l'une des espèces invasives les plus menaçantes au monde (Global Invasive Species Database, http ://issg.org/database). En effet, son invasion représente une menace prise très au sérieux pour la santé des populations humaines du fait de son statut de vecteur avéré pour les virus DENV et CHIKV, et sa compétence [1] pour un grand nombre de virus et pathogènes (1.2.1 - Vecteur de pathogènes).

### L'invasion mondiale

L'histoire de l'invasion d'*Ae. albopictus*, en lien avec des données de génétique des populations sera traitée en détail dans le chapitre suivant. Brièvement, l'aire de répartition "pré-invasion" du moustique tigre inclut le sud et l'est de l'Asie : de l'Inde à l'ouest jusqu'au Japon au nord-est, elle recouvre ainsi les Philippines, la péninsule Indochinoise, Taïwan et la Chine. L'espèce aurait récemment recolonisé ces territoires depuis le Sundaland, qui lors du dernier maximum glaciaire représentait un territoire émergé entre la péninsule indochinoise et l'Indonésie (Porretta *et al.* (2012), cette étude sera détaillée dans le chapitre 2). Les limites de son aire de répartition "originelle" sont cependant floues, en particulier parce que l'invasion mondiale semble avoir débuté peu de temps après les premières descriptions de l'espèce.

Les îles du sud-ouest de l'océan Indien comme Madagascar ou la Réunion, pourraient être les premiers territoires à avoir été colonisés à la faveur des premières migrations humaines en bateau depuis les îles indonésiennes, il y 1000 à 1500 ans ; une présence supposée qui peut avoir été renforcée lors du commerce des épices au XVII$^{ème}$ et XVIII$^{ème}$ siècles (Delatte *et al.* 2011). L'intensification des échanges commerciaux au cours du XX$^{ème}$ siècle a par la suite contribué à la colonisation de l'ensemble des continents, à l'exception de l'Antarctique (Figure 10).

1. Capacité au laboratoire à assimiler, répliquer et transmettre le pathogène

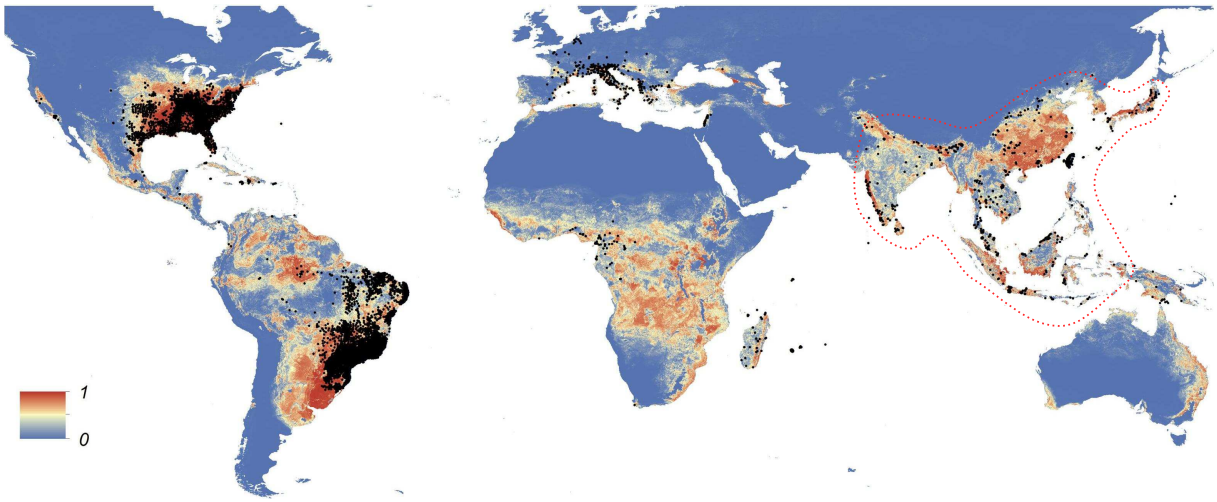



**Figure 10** – Répartition mondiale (points noirs) et projection de l'aire de répartition à partir des données climatiques (dégradé du bleu : 0% de probabilité au rouge 100% de probabilité) du moustique tigre. Entourée en pointillés rouges, l'aire de répartition "avant invasion" estimée. (Modifié d'après Kraemer *et al.* 2015) CC0

En Europe, où il a d'abord été détecté en 1979 en Albanie puis en 1990 en Italie, il est aujourd'hui bien implanté sur la partie nord du pourtour méditerranéen, du sud-est de l'Espagne à la Grèce (Figure 11). Sa présence est avérée depuis 1985 sur le continent américain où son introduction au Brésil et aux États-Unis furent quasi-simultanées (Sprenger et Wuithiranyagool 1986 ; Forattini 1986). Au cours des années 2000, le moustique tigre s'est implanté en Afrique sub-saharienne, notamment au Nigeria, Cameroun, Guinée Équatoriale et Gabon (Paupy *et al.* 2009 ; Bonizzoni *et al.* 2013).

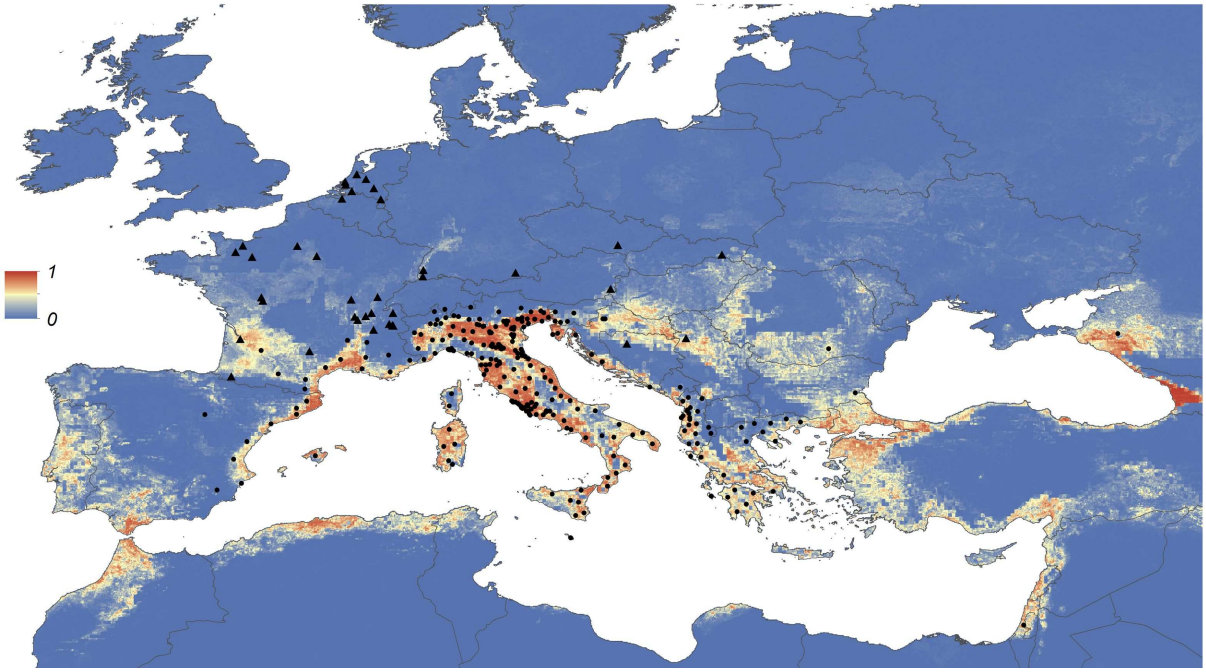



**Figure 11** – Répartition (point noirs : installé, triangles : détecté) et projection de l'aire de répartition du moustique tigre à partir des données climatiques en Europe (dégradé du bleu : 0% de probabilité au rouge 100% de probabilité), (Modifié d'après Kraemer *et al.* 2015) CC0

### Menaces pour l'Homme

Longtemps considéré comme un vecteur secondaire par rapport à *Ae. aegypti*, le moustique tigre a récemment été reconnu responsable, en tant que principal ou seul vecteur, lors d'épidémies de Dengue à Hawaï, en Asie et en Afrique (Bonizzoni *et al.* 2013 ; Tsuda *et al.* 2015).

Le moustique tigre s'est aussi révélé au grand public en tant que principal vecteur du CHIKV lors des importantes épidémies qui ont touché les îles de La Réunion, Maurice, Mayotte et Madagascar entre 2005 et 2007, ainsi qu'en Afrique centrale (de Lamballerie *et al.* 2008 ; Paupy *et al.* 2009 ; Bonizzoni *et al.* 2013).

La part importante prise par *Ae. albopictus* en tant que vecteur du CHIKV semble être notamment due à un cas très préoccupant d'évolution parallèle du virus vers une forme particulièrement adaptée à l'espèce. En effet, trois évènements indépendants de substitution d'un résidu Alanine par une Valine ont eu lieu à la position 226 (A226V) du gène viral *E1*, assurant au virus une meilleure réplication et transmission par le moustique tigre (de Lamballerie *et al.* 2008). Récemment, ce génotype viral a été reconnu responsable d'une épidémie de CHIKV en Thaïlande (Wanlapakorn *et al.* 2014) ; la diffusion de cette nouvelle forme virale dans ce pays démontre que l'invasion globale d'*Ae. albopictus* peut aussi avoir des conséquences néfastes sur les territoires où il est considéré comme endémique.

*Ae. albopictus* est aussi à l'origine de la première épidémie européenne du CHIKV en en Italie en 2007, un territoire où aucun vecteur de ce virus n'était présent auparavant (de Lamballerie *et al.* 2008). D'autre part, plusieurs évènements sporadiques de transmissions autochtone de la dengue et du chikungunya ont été recensés en France et en Europe ces dernières années (l'Institut de veille sanitaire répertorie 5 cas de dengue autochtone en France du premier mai au 21 septembre 2015 ; http ://www.invs.sante.fr) confirmant les risques épidémiologiques liés à la présence du vecteur.

### Les clés du succès : adaptation ou plasticité ?

Le succès invasif du moustique tigre réside dans la diversité des traits écologiques décrits précédemment (1.2.1). L'utilisation de sites de ponte artificiels a notamment contribué à sa dispersion via le commerce maritime des pneus usagés, et des moustiques ont aussi été interceptés dans les contenants de *"lucky bamboo"*, plantes ornementales du genre *Dracaena*, aux États-Unis et aux Pays-Bas provenant de la Chine (Madon *et al.* 2002 ; Schaffner *et al.* 2004). La réussite de cette diffusion par des contenants artificiels est aussi favorisée grâce à la possibilité de ses œufs de résister aux conditions défavorables lors du transport par la quiescence ou la diapause. La diapause est certainement l'un des atouts majeurs pour la colonisation des zones tempérées, permettant aux individus de passer les hivers sous la forme d'un œuf particulièrement résistant. Nous avons aussi vu que le mous-

tique tigre pouvait parfois présenter un avantage compétitif, lui permettant notamment de se maintenir *via* le remplacement ou la cohabitation avec des espèces locales.

Cette amplitude dans la niche écologique d'*Ae. albopictus*, a parfois été qualifiée de "plasticité écologique" (Delatte *et al.* 2008 ; Paupy *et al.* 2009 ; Porretta *et al.* 2012 ; Bonizzoni *et al.* 2013). Cependant, ce terme entretient un flou sur les processus biologiques à l'origine de l'adéquation du moustique aux différents environnements dans lequel il est présent. En effet, les traits favorables à l'invasion du moustique tigre que nous venons de citer correspondent à la compilation des observations faites sur un ensemble de populations, et ne permettent donc pas de distinguer lesquels sont réellement plastiques (c.f. 1.1.1) de ceux qui sont le produit de l'adaptation et donc relèvent d'un polymorphisme génétique au sein de l'espèce.

Des cas de plasticité phénotypiques ont été documentés chez *Ae. albopictus*. La quiescence, évoquée précédemment, est l'exemple le plus marquant de plasticité dans l'arsenal du moustique tigre. Vitek et Livdahl (2009) ont notamment montré que la fréquence des précipitations (simulée par submersion) pouvait moduler le temps avant l'éclosion des œufs au sein d'une lignée de laboratoire. Ainsi des précipitations rares, qui peuvent être interprétées comme un indice du risque de sécheresse prolongée (dépassant vraisemblablement la tolérance de l'espèce), provoquent une éclosion précoce chez *Ae. albopictus*. Par ailleurs, le régime photo-périodique peut aussi induire une réponse plastique : aux États-Unis, plusieurs populations ont montré une augmentation du temps de développement avec la durée du jour (Yee *et al.* 2012). La plasticité est notamment invoquée, car les différentes populations étudiées présentaient la même réponse alors qu'elles ont été échantillonnées le long d'un gradient géographique où le régime photo-périodique est lui-même variable. Il semble indéniable que la plasticité phénotypique de ces traits contribue au potentiel invasif du moustique, en lui permettant de se maintenir au sein d'environnements différents. Cependant, la norme de réaction, qui correspond à l'amplitude des phénotypes possibles à partir du même génotype, peut présenter une variabilité adaptative, autrement dit résulter d'un polymorphisme génétique sélectionné dans des environnements différents (Lande 2015). Une telle évolution adaptative de la plasticité n'a, à notre connaissance, pas encore été recherchée chez *Ae. albopictus*.

Les seuls exemples d'adaptation connus, favorisant l'invasion d'*Ae. albopictus*, sont la diapause saisonnière et la résistance des œufs aux températures négatives, qui sont sélectionnées chez les populations installées dans les environnements tempérés (Hawley 1988 ; Denlinger et Armbruster 2014).

Chez le moustique tigre, la diapause saisonnière est induite, seulement chez des populations qui y sont sensibles, par la réduction de la photopériode à l'approche de l'hiver. C'est la femelle adulte qui perçoit ces signaux et produit alors un œuf plus résistant, disposant notamment d'importantes réserves lipidiques, et dont l'embryon stoppera (ou du moins ralentira fortement) son développement au stade de larve pharate (premier stade)

à l'intérieur du chorion (Denlinger et Armbruster 2014). L'incidence de la diapause, qui correspond à la proportion d'œufs pondus entamant réellement ce développement alternatif, est alors déterminée génétiquement et est corrélée à l'origine géographique des populations. Hawley *et al.* (1987) ont par exemple démontré que soumises à un même régime de jours courts, seules des populations provenant du Japon, de Chine, Corée et des Etats-Unis, situées à plus de 25 ° de latitude nord, étaient capables de pondre des œufs diapausants, alors que des populations tropicales, plus proches de l'équateur (en Asie ou dans l'océan Indien), en sont incapables.

Il en va de même pour l'acclimatation des œufs au froid : l'exposition des œufs à des températures comprises entre 5 ° C et 10 ° C permet de les rendre plus résistants lorsqu'ils sont par la suite soumis à des températures négatives. Cependant, ce traitement n'a pas la même incidence en fonction de l'origine géographique des populations, et contrairement aux populations tempérées, les œufs issus de populations tropicales ne parviennent pas à survire à des températures inférieures à -8 ° C malgré leur acclimatation (Hawley *et al.* 1987 ; Hanson et Craig 1994).

La diapause et l'acclimatation au froid peuvent être considérées comme indépendantes : la première est induite chez l'adulte, la seconde au stade œuf, et l'acclimatation des adultes n'influence pas la résistance au froid de leurs œufs (Hanson et Craig 1994). Cependant, la résistance au froid de l'oeuf d'*Ae. albopictus* est maximisée lorsque ces deux processus sont induits successivement (Hanson et Craig 1994).

Alors que chez les Diptères l'adaptation locale est souvent associée à l'observation de clines le long des latitudes pour des traits comme la taille des ailes, la masse du corps ou le volume des œufs, seule l'incidence de la diapause a été clairement démontrée chez *Ae. albopictus* (Urbanski *et al.* 2012 ; Denlinger et Armbruster 2014). La photopériode critique, c'est à dire la durée de jour à laquelle 50% de œufs induits entrent en diapause, est notamment positivement corrélée à la latitude et à l'altitude chez les populations tempérées d'*Ae. albopictus* (Focks *et al.* 1994).

La diapause peut être aussi cruciale lorsque les limites de résistance des œufs par la quiescence sont atteintes, ce qui peut se produire lors des longs trajets *via* le transport maritime. En effet, comparé à *Ae. aegypti* chez qui la diapause est absente, les œufs non diapausants d'*Ae. albopictus* sont moins résistants aux stress liés à la dessiccation (Sota et Mogi 1992).

De récentes analyses de transcriptomique à haut débit (RNAseq) ont permis d'identifier une partie des mécanismes moléculaires impliqués dans la mise en place et le maintien de la diapause chez *Ae. albopictus* (Poelchau *et al.* 2011, 2013b,c ; Huang *et al.* 2015). Des gènes associés au cytosquelette, aux protéines cuticulaires, au cycle cellulaire ou encore au métabolisme des lipides présentent ainsi de forts différentiels d'expressions entre des individus induits ou non pour la diapause. Un autre gène, *pepck* (*phosphophenol pyruvate carboxykinase*), impliqué dans le passage d'un métabolisme aérobie vers l'anaérobie

apparaît aussi central dans le déroulement de la diapause (Poelchau *et al.* 2011, 2013b). Cependant, si ces mécanismes commencent à être élucidés, le ou les polymorphismes génétiques à l'origine des variations adaptatives au sein des populations demeurent inconnus.

# 4 Objectifs de la thèse

L'importance de la sélection naturelle au cours de l'évolution est une question qui passionne les biologistes, au moins depuis sa formalisation par Darwin puis par son extension à la génétique par Fisher en 1930. Le développement des théories neutralistes, à la fois en génétique (Kimura 1968), mais aussi en écologie (Hubbel, 2001) à offert le cadre théorique permettant de tester formellement l'action de la sélection naturelle et ainsi mieux comprendre la part relative des forces évolutives à l'origine de la diversité du vivant, telle que nous pouvons l'observer aujourd'hui.

Généticiens et écologues se sont notamment consacrés depuis longtemps à l'étude des espèces invasives, qui représentent *"une série d'expériences en évolution [...] potentiellement plus informatives que la plus part des travaux réalisés en laboratoire"* (Waddigton, dans l'introduction de l'ouvrage édité par Backer et Stebbins en 1965 "The genetics of colonizing species"). Par ailleurs, parce que certaines représentent une menace pour la santé ou les activités humaines, mais aussi parce que l'Homme se pose dorénavant la question de son influence sur la biodiversité, les espèces invasives se positionnent à l'interface entre les sciences et la société.

En tant qu'espèce invasive, Le moustique tigre représente le lien qui existe entre les intérêts appliqués, liés ici à la santé publique et de nombreux intérêts fondamentaux, parmi lesquels la génétique de l'adaptation. Ces différents aspects, complémentaires, font de cette espèce un modèle de choix. Si comme nous l'avons évoqué un certain nombre de traits permettent d'expliquer son succès écologique, il existe peu d'informations sur la nature des processus ayant mené à une telle adéquation entre les différentes populations du moustique tigre et leurs environnements respectifs.

Au cours de cette thèse, nous avons donc cherché à déterminer le rôle la sélection naturelle dans l'adéquation du moustique tigre aux divers environnements colonisés. Pour cela, notre objectif principal a été de mener un scan génomique sans *a priori* basé sur la différenciation génétique entre des populations tropicales de l'aire d'origine et des populations invasives en milieu tempéré.

Au préalable, il a été nécessaire de rassembler et synthétiser les informations disponibles concernant la structure génétique des populations du moustique tigre. Ces données sont par exemple nécessaires (en plus des analyses de génétiques des populations réalisées à partir des données collectées) afin de choisir un modèle adapté lors de la réalisation d'un scan génomique (voir 1.1.2). Une telle revue nous informe aussi sur la structure des populations présentes dans l'aire d'origine et les relations qui existent à la fois entre populations natives, invasives, tropicales ou tempérées. Ces données permettent de replacer notre étude de scan génomique dans le contexte global, et de formaliser plus tard des hypothèses quand à l'origine d'évènements d'adaptation ayant eu lieu au sein des populations étudiées. Ce travail de revue bibliographique est le sujet du chapitre 1 et a été

soumis au journal *Heredity*.

Concernant la partie expérimentale, nous disposions de peu de ressources génomiques directement disponibles afin de réaliser une telle expérience. En particulier, le premier assemblage du génome du moustique tigre n'a été publié qu'au mois de septembre 2015. De plus, les marqueurs moléculaires disponibles (voir chapitre 2) n'étaient pas assez nombreux pour envisager leur utilisation dans une expérience de scan génomique.

En conséquence, une grande partie des travaux de cette thèse concernent le développement d'un nouveau marqueur génétique suffisamment dense pour s'acquitter de cette tache. L'exploration conjointe du RAD-seq et du Transposon Display (*C.f.* chapitre 1) nous a conduit à sélectionner la seconde méthode pour la production de ces marqueurs. Nous avons choisi de tirer parti de la large fraction d'ET présente dans le génome d'*Ae. albopictus*, afin de développer une méthode de génotypage basée sur le polymorphisme d'insertion des ET entre les différentes populations étudiées. Des méthodes similaires, mais cependant à plus faible densité de marqueurs, ont été utilisées avec succès chez les moustiques *Anopheles gambiae* (Boulesteix *et al.* 2007 ; Esnault *et al.* 2008) et *Aedes aegypti* (Bonin *et al.* 2008, 2009) respectivement afin de distinguer les formes moléculaires M et S d'*A. gambiae* ou de réaliser un scan génomique pour la résistance à un insecticide.

Le développement du marqueur, à été rendu possible par l'étude du répétome, c'est à dire l'ensemble des séquences répétées d'un génome, d'*Ae. albopictus*. Cette étude, réalisée en collaboration avec Patrick Mavingui et Claire Valiente Moro, qui coordonnent l'un des projet actuel de séquençage du moustique tigre, a permis d'identifier des familles d'ET suffisamment répétées et conservées pour nous permettre de développer le protocole de génotypage au laboratoire. Cette analyse a nécessité le développement d'un outil *ad-hoc*, permettant de détecter, annoter et quantifier les répétitions à partir de lectures (ou *reads*) d'un génome non assemblé, et nous a permis de proposer la première description détaillée du répétome d'*Ae. albopictus* ainsi qu'une analyse comparative avec l'espèce *A. aegypti*. La méthode bioinformatique ainsi que l'analyse génomique ont été publiés en mars 2015 dans la revue *Genome Biology and Evolution* et sont présentés dans le chapitre 3.

Le dernier chapitre est donc l'aboutissement des travaux précédents, et décrit d'une part le protocole original de séquençage à haut débit des insertions d'ET polymorphes entre des populations de l'aire d'origine tropicale (Vietnam) et des populations invasives tempérées (France et Espagne), et le scan génomique d'autre part, basé sur la différenciation de ces marqueurs entre les populations retrouvées dans des environnement différents. Notre hypothèse principale est que la colonisation d'une nouvelle aire bio-géographique par le moustique tigre pourrait être à l'origine de nouvelles pressions de sélections. Si des balayages sélectifs associés au processus adaptatif ont eu lieu, nous devrions être en mesure de détecter une importante différenciation génétique aux polymorphismes voisins des cibles de la sélection.

# Chapitre 1

# Génétique des populations du moustique tigre

*"No I wasn't born in Ghana but Africa is my mama*
*'Cause that's where my mama got her mitochondria"*

– Baba Brinkman, *The rap guide to evolution* 2009

# Avant propos

Les premières études de génétique des populations concernant le moustique tigre ont été réalisées à la fin des années 1980, peu après l'introduction de cette espèce aux États-Unis. Depuis, plusieurs dizaines d'études ont été consacrées à l'analyse de la structure des populations, parfois à l'échelle mondiale, mais aussi à des échelles plus réduites, leur objectif étant le plus souvent d'inférer l'origine des populations invasives.

Dans le cadre de nos travaux, il était important de réaliser une synthèse de ces informations, afin de formaliser nos hypothèses quant au modèle démographique le plus adéquat pour la réalisation du scan génomique. En particulier, nous avions besoin de disposer d'informations sur la structure des populations au sein de l'aire d'origine : il était en effet nécessaire de savoir si les populations que nous comptions étudier au Vietnam étaient ou non représentatives de la diversité génétique présente dans les régions tropicales, et s'il existait une différenciation au sein de l'aire d'origine entre ces populations et celles retrouvées en milieux tempérés. D'autre part, il était aussi nécessaire de faire un point sur l'état de l'art concernant l'origine des populations invasives de par le monde, leur différenciation avec celles présentes dans l'aire d'origine ainsi qu'avec les autres populations invasives, tropicales et tempérées. Encore une fois, ces information devaient permettre de replacer les populations Européennes étudiées dans le contexte général de l'invasion, et évaluer leur représentativité par rapport aux autres populations tempérées. Enfin, d'une manière plus globale, nous souhaitions aussi appréhender les aspects particuliers de la génétique des population de cette espèces et s'il était possible de dresser le portrait d'une population « typique » d'*Ae. albopictus*.

Le résultat marquant de cette synthèse est l'absence globale de structure géographique parmi les populations du moustique tigre. Quelle que soit l'échelle considérée et les marqueurs utilisés, l'essentiel de la variabilité génétique se situe entre les individus au sein des populations, et seule une très faible fraction (souvent moins de 10%) peut être attribuée à une structure continentale. Si les populations invasives tempérées ont pu être reliées aux régions du nord de l'aire d'origine (Japon, Chine), ces résultats restent ténus à la fois au regard de la faible variabilité de certains marqueurs utilisés mais aussi du manque d'exhaustivité concernant l'échantillonnage dans le berceau supposé de l'espèce. Enfin, des résultats récents tendent à indiquer que les populations présentes en Asie du sud-est arborent une très grande diversité génétique sans qu'aucune structure ne puisse être observée, ce qui semble indique que nos populations vietnamiennes peuvent être considérées comme de bonnes représentantes de cette diversité.

Par ailleurs, cette revue dresse un bilan critique des différents marqueurs génétiques disponibles chez *Ae. albopictus*, et présente les résultats de notre expérience préliminaire visant à tester la faisabilité du RAD-seq (méthode décrite dans l'article) chez cette espèce. Ces travaux, réalisés en début de thèse, révèlent que cette méthode se heurte à la com-

plexité du génome d'*Ae. albopictus*. Ces résultats, par ailleurs comparables à ceux obtenus par d'autres chercheurs chez *Aedes aegypti*, nous ont alors encouragés à nous concentrer sur le développement du Transposon Display, initié en parallèle.

# Population genetics of the invasive Asian tiger mosquito *Aedes albopictus*

Clément Goubert[1], Guillaume Minard[2,3], Cristina Vieira[1], and Matthieu Boulesteix[1]

[1]Laboratoire de Biométrie et Biologie Evolutive, UMR CNRS 5558, INRIA, VetAgro Sup, Université Claude Bernard Lyon 1, Villeurbanne, France
[2]Microbial Ecology, UMR CNRS 5557, USC INRA 1364, VetAgro Sup, FR41 BioEnvironment and Health, Université Claude Bernard Lyon 1, Villeurbanne, France
[3]Metapopulation Research Group, Department of Biosciences, University of Helsinki, Helsinki, Finland

**Abstract**

The Asian tiger mosquito *Aedes albopictus* is currently one of the most threatening invasive species in the world. Native to Southeast Asia, the species has spread throughout the world in the last 30 years and is now present in every continent but Antarctica. Because it was the main vector of recent Dengue and Chikungunya outbreaks and because of its competency for numerous other viruses and pathogens, *Ae. albopictus* stands out as a model species for invasive diseases vector studies. A synthesis of the current knowledge about the genetic diversity of *Ae. albopictus* is needed, knowing the interplays between the vector, the pathogens, the environment and their epidemiological consequences. Such resources are also valuable for assessing the role of genetic diversity in the invasive success. We review here the large but sometimes dispersed literature about the population genetics of *Ae. albopictus*. We first present an up-to-date assessment of the available molecular markers and summarize the main genetic characteristics of natural populations. We then synthesize the available data regarding the worldwide structuring of the vector in accordance with its invasion dynamics. Finally, we pinpoint the gaps that remain to be addressed and suggest possible research directions.

**Keywords:** *pathogen vector, invasive species, genetic diversity, genetic structure, molecular markers*

## Introduction

Biological invasions, in spite of their critical impacts on native biodiversity and human societies, represent a special opportunity for population geneticists to observe in naturae phenomena that have otherwise remained mostly theoretical (Bock et al., 2015). For example, the study of a recently installed alien species represents an occasion to investigate the link between genetic diversity, whether neutral or not, and invasion success (Dlugosch and Parker, 2008; Handley et al., 2011).

When such a species is also a threat to human health, as disease vectors are, the collection of empirical knowledge about population dynamics and gene flow can be used to anticipate risks (e.g., modeling spread and epidemiologic implications) and develop control strategies. For example, the existence of multiple locally adapted vector populations could enhance the spread of parasites or arboviruses through space and time (McCoy, 2008). The genetic variation among vector populations also diversifies their interactions with viruses and parasites and alters their transmission dynamics (Barrett et al., 2008). The intensity of gene flow among host populations is also an important point to take into account because it could influence the diffusion of key alleles, such as those involved in insecticide resistance (Caprio and Tabashnik, 1992; Lenormand et al., 1999).

The Asian tiger mosquito *Aedes* (Stegomyia) albopictus (Skuse 1894) has been described as one of the 100 worst invasive species in the world (Global Invasive Species Database, http://www.issg.org/database/). Originating from South and East Asia, this species has spread throughout the world mostly since the second half of the XXth century, and it is now found on every con-

tinent except Antarctica (Kraemer et al., 2015). The new areas colonized by *Ae. albopictus* include such disparate environments as tropical South America and the mostly temperate areas of Northern America and Europe.

Although *Ae. albopictus* has long been considered a vector of secondary importance, its involvement in recent Dengue (DENV) and Chikungunya (CHIKV) outbreaks and its competence for numerous other arboviruses and nematode parasites (Paupy et al., 2009) emphasizes the need to more extensively study the biology of this species (Bonizzoni et al., 2013). *Ae. albopictus* has played an important role in the reemergence of DENV and CHIKV in some of the recently invaded regions (Teixeira et al., 2009; Paupy et al., 2010), and it has also been implicated in new outbreaks, such as the 2007 CHIKV episode in Italy (Rezza et al., 2007). Thus, it represents a relevant case study where invasion genetics could bring substantial improvement in vector survey and management activities.

In addition, knowledge about the genetic diversity and the genetic structure of the Asian tiger mosquito could help researchers evaluate the risk of disease and insecticide resistance spread as well as identify the origins and frequencies of introductions. Such knowledge is also required to search for loci that could be involved in environmental adaptation. The genetic structure of a substantial number of native and invasive populations of this species has been investigated throughout the world, multiplying the number of population genetics reports that are available. They are however highly disparate and range from local reports to worldwide genotype comparisons.

Regarding the increasing interest in such data about *Ae. albopictus*, we present here a comprehensive synthesis of what is known and the questions that remain to be investigated. In the following review, we first exhaustively describe the genetic markers previously developed for *Ae. albopictus*. Then, we try to characterize the key features of the genetic structure of natural populations, and we highlight some characteristics of invasive populations. Finally, we examine different attempts to resolve the geographical origin of invasive populations and suggest directions for future studies of *Ae. albopictus*.

## Genetic markers available in *Ae. albopictus*

Useful genetic markers are expected to have several key features: selective neutrality, ease of scoring in all specimens of the species and sufficient variability to allow for measures of genetic differentiation and genetic clustering of individuals. They also should be robust to re-genotyping and/or allow comparison with genotypes from new samples. Depending on the purpose, the selected markers should be suitable for phylogeography analyses (such as mtDNA), allow interpretation about the breeding structure (such as co-dominant allozymes or microsatellites), or be sufficiently numerous in the genome to tease apart the effects of demographic history from those of natural selection.

**Allozymes**. The first population genetics studies in *Ae. albopictus* were conducted in the late 1980s, using a set of 7 to 10 polymorphic enzymes (Black, Ferrari, et al., 1988; Black, Hawley, et al., 1988; Kambhampati et al., 1991). More recently, Urbanelli et al. (2000) conducted an analysis in Italian populations using a set of 19 enzymatic loci. Such studies allowed for the investigation of breeding structure and fine-scale analysis of the genetic relationships of different populations from around the world. These markers have remarkable resolution, but they were soon withdrawn at the benefit of mitochondrial sequences (mtDNA).

**mtDNA.** The genetic diversity of *Ae. albopictus* has been investigated with three mtDNA genes: cytochrome c oxidase subunit I (*COI*), cytochrome b (*Cytb*) and NADH dehydrogenase subunit 5 (*ND5*). Overall, these loci showed a low level of genetic variation: for example, a study involving a hundred of individuals from around the world revealed only 13 haplotypes using *ND5* (Usmani-Brown, 2009). This low diversity, which first had been attributed to small effective population sizes of the founder populations, was later suggested to be a consequence of a selective sweep induced by *Wolbachia* infection (Armbruster et al., 2003, see also the dedicated Wolbachia paragraph). *COI* and *ND5* were mostly used, either alone or in combination (Birungi and Munstermann, 2002; Mousson et al., 2005; Usmani-Brown, 2009; Žitko et al., 2011; Kamgang et al., 2011; Delatte et al., 2011;

Porretta et al., 2012; Shaikevich and Talbalaghi, 2013; Zhong et al., 2013; Zawani et al., 2014). The ease of use and low cost of these markers could explain why so many studies relied on them at the cost of information loss (especially compared to enzymatic markers). An interesting feature of the mtDNA markers is that the genotype information provided is robust, meaning that a dataset of sequences from other published samples could be added in order to – for example – infer the origin of the population studied (which is difficult to achieve with allozymes or microsatellites). However, for the above-mentioned comparison to be feasible, the sequences available in data banks should be of similar size and from the same region of the gene. That is not the case in *Ae. albopictus*, where a survey of the available sequences in GenBank for *COI* shows that only a small region of 295 bp is shared (Fig. 1). Furthermore, Shaikevich and Talbalaghi (2013) noted that among the sequences available in the data banks, the 5' end of the *COI* sequences displayed more polymorphism than other parts. Accordingly, we would recommend for further analysis based on mtDNA sequences favoring the use of *COI*, using the PCR primers from Zhong et al. (2013), which encompass a large (1,433 bp) sequence that contains the portions used in most of the published studies. This includes the highly polymorphic 5' end, and also a large portion of the 3' end, which only Porretta et al. (2012) had previously amplified and for which the authors found a high level of polymorphism in their Asian samples.

**ITS2.** Recently, a study of ribosomal DNA internal transcribed spacer 2 (ITS2), a marker normally used for inter-specific comparison (Higa et al., 2010), showed interesting results for *Ae. albopictus* (Shaikevich and Talbalaghi, 2013; Manni et al., 2015). Compared to a COI sequence, ITS2 showed 10 time more genetic variation within a sample of 14 north Italian individuals (Shaikevich and Talbalaghi, 2013). Moreover, this region presents two main advantages compared to *COI*: (i) it is located in a variable non-coding region bordered by the two highly conserved 5.8S rRNA and 28S rRNA genes, and (ii) it is a nuclear marker that therefore should not be affected in the case of an mtDNA selective sweep. These promising results should encourage researchers to use this marker for future haplotype analysis. However, to access both alleles of this marker, several cloning and sequencing runs are necessary for each individual.

**Microsatellites.** Variable Nucleotide Tandem Repeats (VNTRs), also known as Simple Sequence Repeats (SSRs), are widely used in population genetics analysis. Different alleles of microsatellites differ in size, so alleles can be easily discriminated using electrophoresis techniques, which is a strong advantage over dominant markers. A total of 53 sets of primers have been used to amplify polymorphic microsatellite loci (2 bp to 5 bp tandem repeats) in natural populations (Porretta et al., 2006; Delatte et al., 2013; Beebe et al., 2013; Manni et al., 2015; Medley et al., 2015). Although microsatellites are marker of choice for population genetics analysis, there are some biases associated with the PCR amplification method, such as (i) better amplification of small alleles caused by a lack of polymerase processivity (Wattier et al., 1998), (ii) mutations at the annealing sites of primers that will prevent amplification of specific alleles also called null alleles (Shaw et al., 1999) and (iii) stuttering that are produced by polymerase slipping and results of additional amplification products with different size from the original alleles (Jones et al., 1987; Van Oosterhout et al., 2004). In addition, the reproducibility and efficiency of certain microsatellites could be questioned, given the fact the most authors developed their own markers for their studies. For French and Vietnamese populations, we do not recommend B6, D2, F3, and alb212 microsatellites (Porreta et al., 2006; Delatte et al., 2013). In our hands, these markers often produced a stuttering like profile (Minard et al., unpublished data). Likewise, we were unsuccessful with the 34-72 microsatellite (Huber et al., 2001; Delatte et al., 2013), probably because of null allele fixation in these populations.

**RAPD.** Random Amplified Polymorphic DNA (RAPD) is a method that was widely used in the 1990's to generate a moderate number of dominant genetic markers without any knowledge of the genome. This method was primarily used to produce a linkage map for the Asian tiger mosquito (Mutebi et al., 1997), and it has only been used in two different phylogeographic studies at a local scale (Ayres and Romão, 2002; Gupta and Preet, 2014). There have been 47 and 141 polymorphic

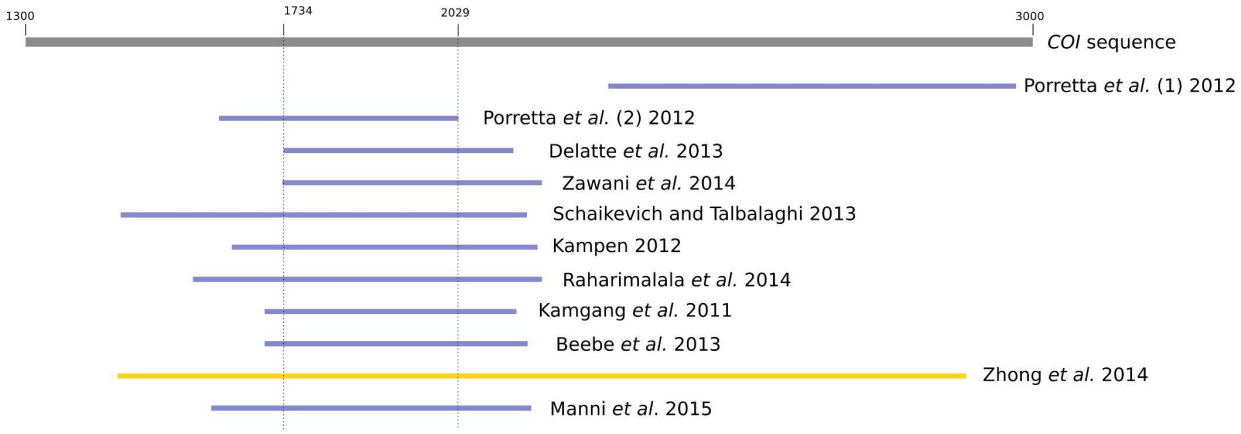



Figure 1: Comparison of the different mtDNA COI regions amplified for phylogeographic studies of *Ae. albopictus* (blue lines). The gray line represents the linear sequence of the COI gene (numbers are the number of base pairs from the origin [NCBI sequence NC_006817.1]). The yellow line is the PCR product amplified by Zhong et al. (2013), showing primers recommended for use.

loci detected in populations with moderate genetic diversity (inferred heterozygosity $0.22 < Hs < 0.35$). However, RAPDs have been shown to be difficult to reproduce in other species (Jones et al., 1997; Perez et al., 1998; Isabel et al., 1999). Thus, the results obtained from these markers should be interpreted with caution.

**RADseq.** Restriction site associated DNA sequencing (RADseq) is currently one of the most popular methods used to study large numbers of SNPs (Single Nucleotide Polymorphisms) in the genomes of non-model species. No study using this method has been published in *Ae. albopictus* yet, but some of us have tested its feasibility in a pilot experiment (CG, CB and MB, unpublished data). Briefly, we used the method described by Henri et al. (2015) to genotype 8 individuals from 2 distinct populations from La Réunion Island, France. We used the SbfI enzyme, which has an 8 bp recognition site and is thus usually considered to be a "low" frequency cutter. Nevertheless, we observed a very large number of sites, producing a total amount of 149 795 RAD loci (using the Stacks pipeline [Catchen et al., 2011]). This number is almost ten times higher than expected by in silico prediction using the genome sequence of the related species *Aedes* aegypti. This high number of loci implies that RADseq, at least in this simple version, cannot be performed for many individuals in most labs due to high cost. Interestingly, a RADseq experiment conducted in *Aedes* aegypti (Brown et al., 2014) also using SbfI discovered a very similar number of RAD loci (184 178) using 128 individuals from around the world. In the end, only 1,504 SNPs were effectively used in this study; this number is enough for most genetic structure analysis purposes but remains limited in the context of genomic scans, especially considering the investment needed to produce these genotypes. Investigation on other restriction enzymes or the development of double-digest RADseq (Peterson et al., 2012) could help us to further decrease the complexity of this genome.

**Transposon display.** The insertion polymorphisms created by transposable elements (TEs) make it possible to obtain a very large number of markers in the genome of *Ae. albopictus*, as it has been demonstrated that some TE families have several thousand well-conserved copies in this species (Goubert et al., 2015). The transposon display (TD) technique involves an enzymatic digestion of the total DNA followed by amplification of insertions of a whole TE family using a single pair of primers. TE insertion polymorphism has successfully been used in several studies involving mosquitoes (Bonin et al., 2008; Esnault et al., 2008). Recently, TD using five TE families led us to amplify more than 100 000 insertion loci in 140 *Ae. albopictus* females from Vietnamese and European populations, and we were able to perform population structure analyses (such as multivariate analysis and AMOVAs) and a genome scan, from which we identified candidate loci potentially

involved in environmental adaptation (Goubert et al. in prep.).

## Population genetics of natural populations

**The "population" issue in *Ae. albopictus*.** First, it is important to notice that the boundaries used to describe "populations" in genetic studies of *Ae. albopictus* are fuzzy. Most often, the term "population" refers to one specific sampling site (such as a point in a city or a GPS coordinate). However, it is not always clear whether this "sampling site" encompasses one or multiple discrete sampling points, such as one or several traps or artificial containers, and if so, the average distances between them and the number of specimens sampled in each (for example, see Urbanelli et al., 2000, Zhong et al., 2013 or Manni et al., 2015). In the following discussion, we will use "sampling site" to refer to the location where a population is sampled (e.g., neighborhood, cemetery, park) and "sampling point" to mean the actual unit of sampling within the sampling site (e.g., a trap or an artificial container). In addition, it is crucial to clearly establish whether the collected samples are eggs, larvae or flying adults. This is important because in *Ae. albopictus*, larvae or eggs collected from the same breeding site are likely to belong to the same progeny. This potential drawback could be easily avoided by collecting only a few (ideally one) individuals per ovitrap and setting multiple traps throughout the sampling site, such as in Delatte et al. (2013). Then, even if adults have been shown to have low dispersal capabilities (200-500 m on average [Liew and Curtis, 2004; Lacroix et al., 2009; Marini et al., 2010]), we could confidently expect that flying adults collected at a sample point are more likely to represent the sampling site diversity than eggs or larvae from the same breeding site. If flying adults are collected, it will also be informative to report whether one or more traps were used (or if aspirator/human sampling was performed at a precise point or in multiple places in the sampling site).

Indeed, a comprehensive description of the sampling strategy is needed for the sake of data interpretation, especially for calculating the inbreeding coefficient ($F_{IS}$) and for partitioning the genetic variance (AMOVA). This subject is discussed in the following paragraph.

**Genetic variation and inbreeding.** With the "population" issue in mind, one of the most common features of genetic studies conducted in *Ae. albopictus* is that the largest part of the observed genetic variation is found at the smallest hierarchical level considered, which has been often referred as "within population" (i.e., within the sampling site; Black, Ferrari, et al., 1988; Black, Hawley, et al., 1988; Kambhampati et al., 1991; Zhong et al., 2013; Gupta and Preet, 2014). The amount of variance attributed to this level, also called "between individuals in population", is remarkably high compared to the upper levels (among groups or among populations in groups) and represents more than half of the total genetic variation, regardless of the marker used (allozymes, mtDNA, microsatellites, RAPD or ITS2) and regardless of the population being invasive or native. Recently, new studies have revealed that what has also been called a "high local differentiation" could be caused by a lack of variation at the intra-individual level ($F_{IS}$, Delatte et al., 2013, Manni et al., 2015). In the AMOVA context (Excoffier, 2004), this level would represent the covariance of alleles of a given locus within individuals within populations. For example, Manni et al., (2015), using ITS2 polymorphism, showed that if the lower hierarchical level was "within population", the genetic variance attributed to that level was 84.78%; when they added the "within individuals" level to the AMOVA, the first level fell to 17.51%, and individual level represented 74.36% of the genetic (co)variance. Equivalent results were found by Delatte et al., (2013) in La Réunion Island, where 80.766% of the variation was estimated to be at the individual level using microsatellites. It is important to highlight that these studies encompass native (Thailand), old (La Réunion) and recent (Italy) invasive populations and thus suggest that *Ae. albopictus* populations could share this pattern globally.

The first studies that investigated the genetic structure of natural populations concluded that such a pattern of variation within the sampling sites was most likely due to high genetic drift accompanying the establishment of the local population (i.e., individuals found in a given sampling site), which implies that such "populations" were founded by small

numbers of adult individuals and had low dispersion rates that restricted gene flow (Black, Ferrari, et al., 1988; Black, Hawley, et al., 1988; Kambhampati et al., 1991). This explanation makes sense in light of recent findings because a high level of genetic covariance within individuals would mean that they could have a high level of inbreeding, which is suggested by their breeding structure.

It is interesting to confront this hypothesis with the results from codominant data, such as allozyme and microsatellite data. For allozymes, most studies reported significant deviation from Hardy-Weinberg Equilibrium (HWE) for several loci and populations toward a heterozygotes deficit (Black, Ferrari, et al., 1988; Kambhampati et al., 1991; Urbanelli et al., 2000; Paupy et al., 2001; Vazeille et al., 2001; De Oliveira, 2003). Using microsatellites, the inbreeding coefficient ($F_{IS}$) has been shown to be positive, ranging from 0.1 to 0.2, supporting a high rate of inbreeding (Delatte et al., 2013; Manni et al., 2015; Minard et al., 2015). The homozygotes excess found with microsatellites has sometimes been attributed to the presence of null alleles (Kamgang et al., 2011 [no $F_{IS}$ data], Delatte et al., 2013), but this should be interpreted with caution because estimation of null allele frequencies is more difficult in populations that are not actually at HWE and if no independent prior of the actual inbreeding level is given (Van Oosterhout et al., 2006; Chybicki and Burczyk, 2009). However, Manni et al. (2015) obtained similar results with new markers with an apparently low level of null alleles. Even if we cannot exclude the possibility that significant $F_{IS}$ values could reflect consequences of a Whalhund effect due to the unknown population structure, the repeated presence of such cues (either high intra-individual genetic co-variance or significant $F_{IS}$) among the studies considered here highly support the inbreeding hypothesis. Thus, a credible portrait of a "typical" population of *Ae. albopictus* would be a network of interconnected breeding sites that each have a high level of inbreeding. Kinship analysis among individuals in sampling points and sampling sites would be very useful to test this hypothesis. Filling the gap between genetic structure and knowledge from behavioral ecology would also be extremely valuable, especially regarding the mating behavior of *Ae. albopictus*. Such studies would help to define the boundaries of what should be considered a "population" and could suggest a uniform sampling strategy for use for future studies. Meanwhile, as remarked earlier, it is important to carefully report the sampling strategy, especially the distance between breeding sites if sites are used as the sampling unit.

**Urban structuring.** One feature of *Ae. albopictus* is its preference for peri-urban areas, where both breeding sites and hosts are available. Urbanization level has been suggested to limit the migration rates between sampling sites. Indeed, Vazeille et al. (2001) showed in Madagascar that sampling sites from Antsiranana (a large community of 105 000 inhabitants) were much more differentiated (N=4, FST=0.227) than samples from Joffreville (N=3, FST=0.095; a small rural town of 5 000 inhabitants). However, no information regarding the distance between sampling sites inside each city is available. In La Réunion Island, populations from the west coast (dryer and much more populated than the east coast) showed the highest differentiation level (Paupy et al., 2001). If further studies are encouraged to confirm this pattern, it can be suggested that the numerous but scattered suitable breeding sites in highly urban environments combined with the low dispersal of the mosquito could contribute to higher differentiation of local populations than in rural or natural areas.

***Wolbachia.*** The proteobacterium *Wolbachia* is a cytoplasmic symbiont that has been found in more than 40% of arthropod species (Werren et al., 1995; Zug and Hammerstein, 2012). This symbiont is often involved in reproductive parasitic interaction with its arthropod hosts, and it can therefore influence the genetic structure of populations (reviewed by Charlat et al., 2003; Hurst and Jiggins, 2005). The Asian tiger mosquito is infected by two Wolbachia strains named *w*AlbA and *w*AlbB (Sinkins et al., 1995; Zhou et al., 1998). Their prevalence in natural populations was estimated to be nearly 100%, and they are mostly found simultaneously as a super-infection inside their host (Werren et al., 1995; Kittayapong et al., 2002; de Albuquerque et al., 2011; Zouache et al., 2011; Bourtzis et al., 2014; Minard et al., 2015). These bacteria control mosquito reproduction by inducing bidirectional cytoplasmic incompatibility. This phenotype prevents infected males from generating progeny when they breed with an uninfected female

or a female infected with a different Wolbachia strain. However, super-infected females can produce viable progeny regardless of the infection status of the males and thus have increased fitness for reproduction (Dobson et al., 2004). Probably because of natural selection, super-infected females are highly prevalent in most populations that can therefore be considered panmictic. Interestingly, the ftsz and wsp genes from both *Wolbachia* clades do not show any variation across geographically distant populations (Armbruster et al., 2003; de Albuquerque et al., 2011). This observation suggests a recent invasion of the symbionts in mosquito populations. Such a recent invasion can drastically modify the structure of mitochondrial genes. Indeed, because the symbiont and mitochondria are transmitted together to the progeny, linkage disequilibrium can occur between them when the symbiont spreads (Hurst and Jiggins, 2005). Such mitochondrial hitchhiking has been demonstrated with laboratory populations of *Ae. albopictus* (Kambhampati and Verleye, 1992). The authors observed fixation of a mitochondrial variant after only two generations with unidirectional cytoplasmic incompatibility. In natural populations, such mitochondrial selection has also been suggested. Indeed most of mitochondrial markers show extremely low diversity (Kambhampati et al., 1991; Maia et al., 2009; Usmani-Brown, 2009; Kamgang et al., 2011). However, because Wolbachia invasion in *Ae. albopictus* has never been dated and because several parts of the mtDNA remain variable (e.g., the 5' end of the COI gene), the impact of Wolbachia mitochondrial hitchhiking has never been clearly demonstrated.

**Invasive populations.** It is often assumed that invasive populations are subject to demographic events, such as bottlenecks, that could increase the importance of genetic drift and lead to reduced genetic diversity in the invaded range (Handley et al., 2011). However, the influence of such events on genetic diversity is modulated and sometimes counterbalanced by several factors, such as the amount of invaders and the frequency of introduction (i.e., the propagule pressure) (Handley et al., 2011; Bock et al., 2015). In *Ae. albopictus*, few studies have specifically focused on these aspects, and there is no evidence that invasions have been followed by a reduction of genetic variability. On the contrary, there are several indications of repeated and possibly massive introductions (detailed in the next chapter). In addition, the colonization pattern of *Ae. albopictus* is characterized by an absence of a natural progressive wave front (that could be accompanied by allele "surfing" at the edge of the invasive range) but is well explained by human-mediated dissemination via the transportation infrastructure (Medlock et al., 2012; Roche et al., 2015).

## Worldwide genetic structure of native and invasive populations

**Genetic diversity in the native area.** To determine the origins of invasive populations, it is important to look at both genetic diversity and structure in the native area. This ranges from Southern (including India) and tropical Eastern Asia (incl. Indonesia) to East China and Japan (Hawley, 1988; Bonizzoni et al., 2013). One weakness of several studies looking at the origins of invasive populations is the use of a restricted sample from the native range (Birungi and Munstermann, 2002; Mousson et al., 2005; Usmani-Brown, 2009) or even the use of laboratory samples that have low genetic variation (Birungi and Munstermann, 2002). Second, most of these studies used markers with low variation such as certain mtDNA sequences (Birungi and Munstermann, 2002; Mousson et al., 2005; Usmani-Brown, 2009; Haddad et al., 2012; Shaikevich and Talbalaghi, 2013; Zawani et al., 2014). Recently, and since the development of new COI primers, two studies offered a more comprehensive view of the genetic diversity in native areas. Porreta et al. (2012) and Zhong et al. (2013) showed that the expected native area, including Japan, shows high genetic diversity (62 haplotypes found in 174 individuals, and 66/346, respectively) with little or no genetic structure. Combining phylogeography and species distribution modeling (using climate data), Porroreta et al. (2012) suggest that the current distribution of *Ae. albopictus* in SE continental Asia, Japan and the Indochinese peninsula is the result of recolonization from a single but large and ecologically diverse area, the Sundaland, which existed during the Last Glacial Maximum (LGM, 21 000 years BP) as an area connecting Sumatra, Java and Borneo to the Indochinese peninsula. The authors argue that a single progressive range expansion, from Sundaland to the Northern territories could have begun before the

LGM ( 70 000 years BP) and continued until recent times. In addition, the northernmost populations studied displayed less genetic diversity and more derived haplotypes than the southernmost ones, which is consistent with a more recent expansion event, facilitated by a switch from hunting-gathering to farming in human populations. Humans also expanded their range from south to north in Asia 15 000 years ago. Thus, the lack of genetic structure in this part of the native area could be explained by conservation of great diversity due to the presence of contrasted but interconnected ecosystems in a unique southern refuge, followed by one recent recolonization event. Even if no clear genetic structure has been reported, the above-mentioned studies did not include populations from the southern range of *Ae. albopictus* (such as Sumatra, Java or Borneo), with the exception of Zhong et al. (2014), who included a single sampling site from Singapore. They found that while Chinese, Taiwanese and Japanese populations showed a very low level of differentiation, they were all well differentiated from the Singapore population, suggesting a possible structuring between northern and southern insular populations. Thus, it would be interesting to investigate the genetic diversity and structure of these populations and compare them to continental ones. Furthermore, it is worth mentioning that allozyme studies have shown evidence for differentiation between the southern insular populations and northern ones, as well as between western (India, Sri Lanka) and eastern populations of the native area (Kambhampati et al., 1991; Urbanelli et al., 2000; see also Fig. 2). As long as the absence of structure in the whole native area is not fully demonstrated, a complete sampling to infer the origin of native populations should include samples from all areas (continental SE Asia, insular SE Asia, the Japanese peninsula, and a western location such as India).

**South-West islands of the Indian Ocean (SWIO) [ 1500 y. BP to -].** SWIO includes Madagascar, La Réunion and other smaller islands such as Mayotte, Mauritius, Rodrigue, Glorieuse Comoros and Seychelles. Delatte et al. (2011) identified two distinct genetic groups in this area using a COI polymorphism. The first group includes numerous widespread haplotypes from samplings performed in 2006-2007, and the second contains much older samples (1956 and 1992) from Madagascar and La Réunion. Based on anthropological research, Delatte et al. (2011) hypothesized that colonization of SWIO by *Ae. albopictus* could have begun 1 500-2 000 years ago with the arrival in Madagascar of humans from Indonesia; several human colonizations occurred thereafter, and SWIO were also frequented by spice traders in the 17th-18th centuries, probably favoring the spread of *Ae. albopictus* in this area. According to the authors, the older genetic group could represent this original colonization of *Ae. albopictus*, whereas the more recent and widespread group could represent the second, modern wave of invasion. However, the geographical origin of this recent invasion was not investigated in this study. Using a combination of the mtDNA markers previously mentioned, nine population samples from Madagascar and La Réunion were only slightly differentiated from four samples of the Indochinese peninsula and Brazil, but there was little statistical support because there were only four informative sites in the mtDNA alignment (Mousson et al., 2005; Raharimalala et al., 2012). Earlier, Kambampathi et al. (1991) showed that 24 individuals from a single sampling site collected in 1988 in Madagascar and genotyped at enzymatic loci formed a distinct cluster from samples of the native area, including Sri Lanka, China, Japan, India, Borneo and Malaysia (see Fig. 2A); although it only represents one "population", this sample could belong to the older genetic group described by Delatte et al. (2011), which would have been significantly differentiated from native populations in SE Asia at that time.

**Hawaii [ 19th century -].** The Hawaiian islands are a location of interest in the biogeography of *Ae. albopictus* because they represent a possible source of introduction for trans-Pacific territories, especially the USA. Zhong et al. (2013) showed that the genetic diversity found in Hawaii is very similar to that of samples from continental SE Asia, but is well differentiated fromthat found in Singapore. Usmani-Brown et al. (2009) found the highest haplotype diversity in their worldwide survey in a ND5 mtDNA fragment in Hawaii, and they hypothesized according to the low variability of this marker elsewhere that such a pattern should be the results of a longer presence time of older invasive populations in Hawaii compared to the other introduced populations. In addition, Mousson et al. (2005) found that the strain isolated in Hawaii

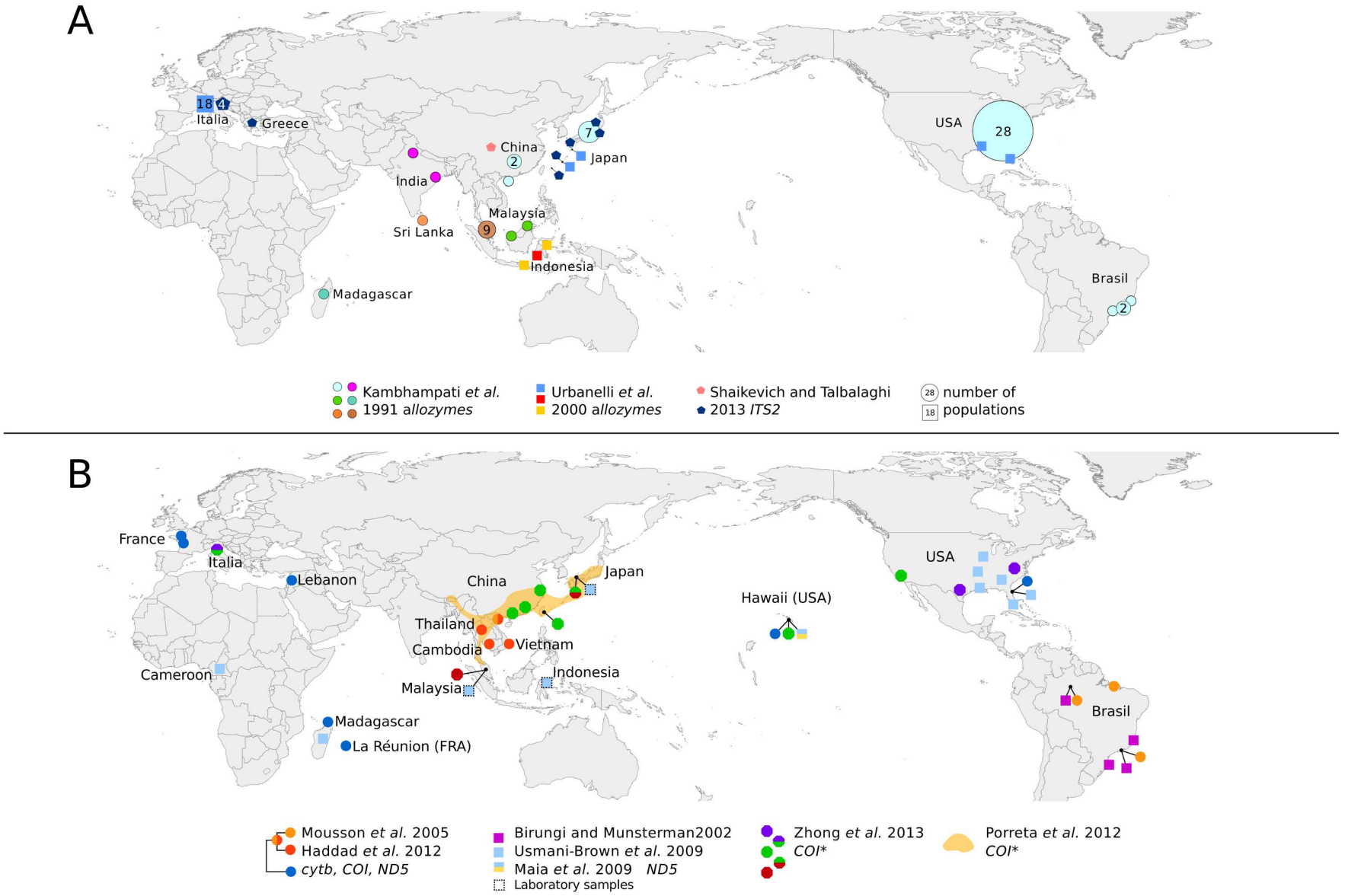



Figure 2: World maps representing homogeneous genetic groups of *Ae. albopictus* identified using **A.** nuclear markers or **B.** mtDNA markers. The results of comparable studies are represented with the same shape. Colors indicate genetic groups, and the types of markers are italicized. Mixed colors for mtDNA markers indicate either intermediate haplotypes (Mousson et al., 2005) or mixed sampling sites. In B, the orange-colored area represents the sampling of Porreta et al. (2012), who found no genetic structure. Asterisks refer to the new highly polymorphic *COI* sequences.

was differentiated from that of the Indochinese peninsula (see Fig. 2B); however, with no other representatives of the native area, few conclusions could be drawn.

**The Americas [1985-].** Colonization of the Americas, and especially the USA, by *Ae. albopictus* in the 1980s was the starting point for studies of the population genetics of this species. The first invasive populations settled in Texas, and the species was quickly found in many important cities of the Southwest and Midwest United States (Sprenger and Wuithiranyagool, 1986; Black, Ferrari, et al., 1988; Black, Hawley, et al., 1988). At the same time, *Ae. albopictus* was also found in Brazil (Forattini, 1986). An early enzymatic survey related the Brazilian samples to Japanese and Chinese ones (Kambhampati et al., 1991), which surprised the authors because Brazilian individuals, which do not respond to photoperiodic diapause induction, were expected to be genetically distinct from US and Japanese ones, for which seasonal diapause can be triggered. However, using mtDNA polymorphism, Birungi and Munsterman (2002) showed that one mutation step discriminates US and Brazilian populations at the ND5 marker, and the marker later showed no variation between *Ae. albopictus* populations anywhere in the world except Brazil (Maia et al., 2009; Usmani-Brown, 2009). Mousson et al. (2005) also found evidence that Brazilian populations could belong to a separate genetic group because they formed a slightly different phylogenetic group with populations of the Indochinese peninsula at mtDNA markers (Cytb-ND5-COI, see Fig. 2B). However, these results are tenuous, and further investigations about the origin of Brazilian populations are needed. In the USA, surveys suggested that a large number of individuals were repeatedly introduced due to intense maritime exchanges, which allowed *Ae. albopictus* to settle (Reiter and Sprenger, 1987; Moore, 1999; Zhong et al., 2013). For example, early after the first reports of invasion, Kambhampati et al. (1990) performed a five year survey of allele frequencies in several localities in the country, and they found no reduction of heterozygosity, with occasional increases in certain cities and a rapid increase in effective population sizes. Most of the surveys acknowledged that US populations were mostly related to the northernmost Asian populations, including Japan (Kambhampati et al., 1991; Urbanelli et al., 2000; Birungi and Munsterman 2002, Mousson et al., 2005, Vargas et al., 2013). However, Zhong et al. (2013) showed a clear distinction between two Eastern (Texas and New Jersey) populations and several samples from Los Angeles (LA) that settled only recently in spite of a high propagule pressure. LA individuals were mostly associated with the northernmost Asian samples (China, Taiwan and Japan), and individuals in formerly colonized Eastern US cities were mostly assigned to a genetic cluster not found in Asia (see Fig. 2). Interestingly, some individuals sampled in Los Angeles in 2001 were related to Singapore haplotypes and were not found in 2011, suggesting that only genotypes from temperate climates could have settled successfully in the USA.

**Europe [1979-].** The first invasive population was reported in 1979 in Albania (Adhami and Murati, 1987), but the spread of the Asian tiger mosquito in this continent seems to have dramatically increased since the 1990s. Italy is probably the country of most concern because of the high density of mosquito populations, which contributed to the 2007 Chikungunya outbreak. Surveys using various types of genetic markers link the Italian samples to Japanese and US ones (Urbanelli et al., 2000 [allozymes], Zhong et al., 2013 [COI], Shaikevich and Talbalaghi 2013 [ITS2]), but this picture could be biased: in two of these studies, the group from which which Italian populations were discriminated are either represented by one population (Zhong et al., 2013) or one genotype (Shaikevich and Talbalaghi 2013). A recent survey by Manni et al. (2015) using microsatellites and ITS2 polymorphism compared populations from Italy, La Réunion Island and Thailand; it revealed that in spite of a high polymorphism found with those markers, no continental structuring of the populations could be found. For example, some Italian populations were as different from each other as the Thailand and La Réunion populations, and the phylogeny of ITS2 haplotypes revealed no structure at all (however, that is a common pattern in *Ae. albopictus*, see "Genetic variation and inbreeding" in the first part of this review). Recent work comparing French and Vietnamese populations showed that a significant but small difference exists between the continents (Minard et al.,

2015), but we also found that a Spanish population (near Barcelona) was as much differenciated from the Vietnamese as from the French populations (Transposon Display, Goubert et al. in prep.). Once again, because the aim of those studies was not to find the origin of the invasive populations, they did not include a sufficient number of populations from the native area. Consequently, those findings are not strong enough to strongly reject structure between continents. Finally, the most informative study could be the older one, where Urbanelli et al., (2000), using allozymes, found evidence for structure between populations from Italy, US and Japan vs populations from several Indonesian islands (see Fig. 2A). This underscores the need to include both North and South-East Asian populations. Due to climatic similarities, it seems important to also compare European populations with those in northern and continental locations of the native range, such as China, that have been far less studied but could be a credible source of European invaders.

**Africa.** In central Africa, *Ae. albopictus* has been reported in Cameroon, Nigeria, Equatorial Guinea, Gabon and Central Africa Republic (Paupy et al., 2009; Kamgang et al., 2013; Bonizzoni et al., 2013). The invasive populations of *Ae. albopictus* have been studied in Cameroon using COI and microsatellites (Kamgang et al., 2011). The identification of two distinct genetic clusters, present in the same locations (microsatellites), led the authors to suspect multiple invasion events. In addition, COI analyses revealed a stronger identity of the Cameroonian haplotypes with tropical populations (India, Thailand, Vietnam, Brazil) than with temperate (France, Greece, USA) or Hawaii and SWIO sources. This contrasts with the findings of Usmani-Brown (2009), who found at the ND5 loci that the individuals sampled in Cameroon were mostly related to several US, Hawaii, Rome (Italy) and SWIO populations. Furthermore, a recent study in the Central African Republic highlighted the relatedness of samples with both the tropical and temperate haplotypes cited earlier (Kamgang et al., 2013). These results suggest multiple sources of introductions for Western-African populations.

**Other locations.** Based on the same combination of mtDNA markers used by Mousson et al. (2005), the geographical origin of individuals found in Lebanon (Haddad et al., 2012) was attributed to temperate areas (see Fig. 2B). However, the low variation of the mtDNA markers used suggests that these results should be interpreted with caution. The geographic origin of Australian invaders, found in the Torres Straits Islands region (North Australia) was investigated using microsatellites, and it was found that this region was not settled from nearby Papua New Guinea (as previously thought) but rather by individuals related to Timor Leste and Jakarta, highlighting the role of humans (probably illegal fishing) in the spread of *Ae. albopictus* (Beebe et al., 2013).

## Conclusions and suggested research directions

This compilation of 30 years of research illuminates the key features of the population genetics of *Ae. albopictus*. The most striking results are the apparent lack of geographical genetic structure according to geography and the high variability found within sampling sites, independently of whether the populations are native or invasive. The lack of isolation by distance seems to be due to a combination of low natural dispersion capabilities and a high level of human mediated spread, as recently demonstrated by Medley et al. (2014) using a landscape genomics framework. The "populations" of *Ae. albopictus* exhibit high levels of inbreeding, which could lead to high contrasts between genotypes from different sampling points in the same site. To validate this hypothesis, more care should be taken to report the sampling strategies, and kinship analysis should be undertaken to study the genetic relationships within and between the sampling points of a sampling site. This will make it possible to understand the original genetic structure that is also found in A. aegypti, the yellow fever mosquito, which is also a mostly human-dispersed and invasive mosquito (Gonçalves da Silva et al., 2012; Damal et al., 2013). However, in A. aegypti, probably because its spread from Africa to America and then Asia is older than the worldwide spread of *Ae. albopictus*, there is genetic evidence for a progressive invasion along with a reduction of genetic diversity from the first to the last colonized continents (Powell and Tabachnick, 2013). This has, to the best of our knowledge, not been shown in *Ae. albopictus*.

Analysis of the available data stresses that future population genetics studies in *Ae. albopictus* should include a more exhaustive worldwide sampling from temperate (including China), sub-tropical and tropical native areas in order to infer the origin of invasive populations. It is important in particular to know whether restricted gene flow really exists between tropical non-diapausing and temperate-diapausing populations in the native area, and if phenotypically similar populations share a genetic kinship. Indeed, the apparent lack of genetic differentiation at a worldwide scale contrasts with the ecological polymorphism found in this species. Particularly, the photoperiodical diapause, which has a demonstrated genetic basis (Urbanski et al., 2012; Poelchau et al., 2013), seems to be an important component of climatic adaptation, favoring the invasive success of *Ae. albopictus*. Thus, the remaining questions in this field include how much of the genome has a low level of differentiation and is it possible to find highly differentiated regions that could be involved in this adaptation? Are there other traits involved in the local adaptation for which we can identify a genetic basis? To what extent could phenotypic plasticity be responsible for both ecological polymorphism and genetic homogeneity? The recent availability of more informative markers and large collections of samples including different time periods is important for the investigation and modeling of invasion routes (Cristescu, 2015). In particular, microsatellites are very useful for determining the genetic relationships of recently introduced populations (Kamgang et al., 2011; Beebe et al., 2013), and the polymorphic COI region should be used for future phylogeographical studies.

Knowledge about the genetic structure will also be valuable for the study of vector-pathogen interactions. Indeed, the genetic variability of vectors would affect the outcomes of close relationships with viruses or pathogens. In A. aegypti, the host genotype modulates the transcriptional response during infection with a strain of DENV virus (Behura et al., 2014), and the way the mosquitoes can acquire the virus when it is present at low levels during feeding (Pongsiri et al., 2014). For *Ae. albopictus*, De Oliveira et al. (2003) and Fernández et al. (2004) showed that the location of the strain (from USA, Caiman Island or Brazil) does not affect the competence for DENV or Yellow Fever Virus. Artificially induced inbreeding depression did not affect infection rate by Plasmodium gallinaceum avian pathogen (O'Donnell and Armbruster, 2010). Recently, Zouache et al. (2014) demonstrated a three-level Genotype X Genotype X Environment (GxGxE) interaction between *Ae. albopictus* strain, CHKV strain and environment, suggesting the need for research in this field. We suggest that future studies should involve multiple strains per geographic location because most of the genetic variability in *Ae. albopictus* is observed at a local scale. It will also be important to take into account the fact that this is one of the insects with the largest genome size (Gregory, 2015; Animal Genome Size Database. http://www.genomesize.com), a large amount of repetitive DNA (Goubert et al., 2015) and potentially a high level of genome size variation (Rao and Rai, 1987; Kumar and Rai, 1990). These patterns could potentially affect the way populations are able to adapt or exchange genes.

The case of *Ae. albopictus* represents a concrete example of a fast and successful invasion, sustained by high propagule pressure and high genetic diversity. Its expansion into the invaded areas is strongly driven by human activities, which are thus actively involved in the shape of the current genetic structure. Because of its epidemiological importance, and also because of its status as an invasive species, the Asian tiger mosquito should be considered a model species for which an increase of knowledge would benefit a large community of researchers. Through the example of *Ae. albopictus*, we show here that knowledge of essential features about the natural populations (breeding structure, dispersal capabilities, local genetic diversity) is useful for tackling more complex problems such as resolving invasion routes, estimating the role of adaptation in biological invasions and predicting the outcomes of host-pathogen interactions. Standardization of sampling and genotyping methods is also recommended in order to avoid the dispersion of data that would become irrelevant for global inference.

## Acknowledgements

We are grateful to Marie Fablet, who provided insightful comments on the manuscript. We also thank Céline Toty (MIVEGEC Montpellier) for providing the *Ae. albopictus* individuals from La Réunion island used in the pilot RAD-seq experiment. C.G. received a grant from the French Ministry of Supe-

rior Education. This work was funded in part by the IUF and the CNRS.

## Conflict of interest

The authors declare no conflict of interest

## References

Adhami J, Murati N (1987). The presence of the mosquito *Aedes albopictus* in Albania. Rev Mjekësore: 13–16.

De Albuquerque AL, Magalhães T, Ayres CFJ (2011). High prevalence and lack of diversity of Wolbachia pipientis in *Aedes albopictus* populations from Northeast Brazil. Mem Inst Oswaldo Cruz 106: 773–776.

Armbruster P, Damsky WE, Giordano R, Birungi J, Munstermann LE, Conn JE (2003). Infection of New- and Old-World *Aedes albopictus* (Diptera: Culicidae) by the Intracellular Parasite Wolbachia: Implications for Host Mitochondrial DNA Evolution. J Med Entomol 40: 356–360.

Ayres CFJ, Romão TPA (2002). Genetic diversity in Brazilian populations of *Aedes albopictus*. Mem Inst Oswaldo Cruz 97: 871–875.

Barrett LG, Thrall PH, Burdon JJ, Linde CC (2008). Life history determines genetic structure and evolutionary potential of host-parasite interactions. Trends Ecol Evol 23: 678–685.

Beebe NW, Ambrose L, Hill L a., Davis JB, Hapgood G, Cooper RD, et al. (2013). Tracing the Tiger: Population Genetics Provides Valuable Insights into the *Aedes* (*Stegomyia*) *albopictus* Invasion of the Australasian Region. PLoS Negl Trop Dis 7: e2361.

Behura SK, Gomez-Machorro C, deBruyn B, Lovin DD, Harker BW, Romero-Severson J, et al. (2014). Influence of mosquito genotype on transcriptional response to dengue virus infection. Funct Integr Genomics 14: 581–9.

Birungi J, Munstermann LE (2002). Genetic Structure of *Aedes albopictus* (Diptera: Culicidae) Populations Based on Mitochondrial *ND5* Sequences: Evidence for an Independent Invasion into Brazil and United States. Ann Entomol Soc Am 95: 125–132.

Black WC, Ferrari JA, Sprengert D (1988). Breeding structure of a colonising species: *Aedes albopictus* (Skuse) in the United States. Heredity (Edinb) 60: 173–181.

Black WC, Hawley WA, Rai KS, Craig GB (1988). Breeding structure of a colonizing species: *Aedes albopictus* (Skuse) in peninsular Malaysia and Borneo. Heredity (Edinb) 61: 439–446.

Bock DG, Caseys C, Cousens RD, Hahn MA, Heredia SM, Hübner S, et al. (2015). What we still don't know about invasion genetics. Mol Ecol 24: 2277–2297.

Bonin A, Paris M, Després L, Tetreau G, David J-P, Kilian A (2008). A MITE-based genotyping method to reveal hundreds of DNA polymorphisms in an animal genome after a few generations of artificial selection. BMC Genomics 9: 459.

Bonizzoni M, Gasperi G, Chen X, James AA (2013). The invasive mosquito species *Aedes albopictus*: current knowledge and future perspectives. Trends Parasitol 29: 460–468.

Boulesteix M, Simard F, Antonio-Nkondjio C, Awono-Ambene HP, Fontenille D, Biémont C (2007). Insertion polymorphism of transposable elements and population structure of Anopheles gambiae M and S molecular forms in Cameroon. Mol Ecol 16: 441–452.

Bourtzis K, Dobson SL, Xi Z, Rasgon JL, Calvitti M, Moreira LA, et al. (2014). Harnessing mosquito-Wolbachia symbiosis for vector and disease control. Acta Trop 132 Suppl: S150–63.

Brown JE, Evans BR, Zheng W, Obas V, Barrera-Martinez L, Egizi A, et al. (2014). Human impacts have shaped historical and recent evolution in *Aedes aegypti*, the dengue and yellow fever mosquito. Evolution 68: 514–525.

Caprio MA, Tabashnik BE (1992). Gene Flow Accelerates Local Adaptation Among Finite Populations: Simulating the Evolution of Insecticide Resistance. J Econ Entomol 85: 611–620.

Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH (2011). Stacks: building and genotyping Loci de novo from short-read sequences. G3 (Bethesda) 1: 171–182.

Charlat S, Hurst GDD, Merçot H (2003). Evolutionary consequences of Wolbachia infections. Trends Genet 19: 217–223.

Chybicki IJ, Burczyk J (2009). Simultaneous estimation of null alleles and inbreeding coefficients. J Hered 100: 106–13.

Cristescu ME (2015). Genetic reconstructions of invasion history. Mol Ecol 24: 2212–25.

Damal K, Murrell EG, Juliano SA, Conn JE, Loew SS (2013). Phylogeography of *Aedes aegypti* (yellow fever mosquito) in South Florida: mtDNA evidence for human-aided dispersal. Am J Trop Med Hyg 89: 482–8.

Delatte H, Bagny L, Brengue C, Bouetard a, Paupy C, Fontenille D (2011). The invaders: phylogeography of dengue and chikungunya viruses Aedes vectors, on the South West islands of the Indian Ocean. Infect Genet Evol 11: 1769–1781.

Delatte H, Toty C, Boyer S, Bouetard A, Bastien F, Fontenille D (2013). Evidence of Habitat Structuring *Aedes albopictus* Populations in Réunion Island. PLoS Negl Trop Dis 7: e2111.

Dlugosch KM, Parker IM (2008). Founding events in species invasions: genetic variation, adaptive evolution, and the role of multiple introductions. Mol Ecol 17: 431–449.

Dobson SL, Rattanadechakul W, Marsland EJ (2004). Fitness advantage and cytoplasmic incompatibility in Wolbachia single- and superinfected *Aedes albopictus*. Heredity (Edinb) 93: 135–142.

Esnault C, Boulesteix M, Duchemin JB, Koffi AA, Chandre F, et al. (2008) High Genetic Differentiation between the M and S Molecular Forms of Anopheles gambiae in Africa. PLoS ONE 3: e1968.

Excoffier L (2004). Analysis of Population Subdivision. In: Handbook of Statistical Genetics, John Wiley & Sons, Ltd: Chichester.

Fernández Z, Moncayo A, Forattini OP, Weaver SC (2004). Susceptibility of Urban and Rural Populations of *Aedes albopictus* from São Paulo State, Brazil, to Infection by Dengue-1 and -2 Viruses. J Med Entomol 41: 961–964.

Forattini OP (1986). Identificação de Aedes (Stegomyia) albopictus (Skuse) no Brasil. Rev Saude Publica 20: 244–245.

Gonçalves da Silva A, Cunha ICL, Santos WS, Luz SLB, Ribolla PEM, Abad-Franch F (2012). Gene flow networks

among American *Aedes aegypti* populations. Evol Appl 5: 664–676.

Goubert C, Modolo L, Vieira C, Valiente-Moro C, Mavingui P, Boulesteix M (2015). De novo assembly and annotation of the Asian tiger mosquito (*Aedes albopictus*) repeatome with dnaPipeTE from raw genomic reads and comparative analysis with the yellow fever mosquito (*Aedes aegypti*). Genome Biol Evol 7: 1192–205.

Gupta S, Preet S (2014). Genetic differentiation of invasive *Aedes albopictus* by RAPD-PCR: Implications for effective vector control. Parasitol Res 113: 2137–2142.

Haddad N, Mousson L, Vazeille M, Chamat S, Tayeh J, Osta MA, et al. (2012). *Aedes albopictus* in Lebanon, a potential risk of arboviruses outbreak. BMC Infect Dis 12: 300.

Handley L-J, Estoup A, Evans DM, Thomas CE, Lombaert E, Facon B, et al. (2011). Ecological genetics of invasive alien species. BioControl 56: 409–428.

Hawley WA (1988). The biology of *Aedes albopictus*. J Am Mosq Control Assoc Suppl 1: 1–39.

Henri H, Cariou M, Terraz G, Martinez S, El Filali A, Veyssiere M, et al. (2015). Optimization of multiplexed RADseq libraries using low-cost adaptors. Genetica 143: 139–43.

Higa Y, Toma T, Tsuda Y, Miyagi I (2010). A multiplex PCR-based molecular identification of five morphologically related, medically important subgenus Stegomyia mosquitoes from the genus Aedes (Diptera: Culicidae) found in the Ryukyu Archipelago, Japan. Jpn J Infect Dis 63: 312–316.

Huber K, Mousson L, Rodhain F, Failloux A-B (2001). Isolation and variability of polymorphic microsatellite loci in *Aedes aegypti*, the vector of dengue viruses. Mol Ecol Notes 1: 219–222.

Hurst GDD, Jiggins FM (2005). Problems with mitochondrial DNA as a marker in population, phylogeographic and phylogenetic studies: the effects of inherited symbionts. Proc Biol Sci 272: 1525–1534.

Isabel N, Beaulieu J, Thériault P, Bousquet J (1999). Direct evidence for biased gene diversity estimates from dominant random amplified polymorphic DNA (RAPD) fingerprints. Mol Ecol 8: 477–483.

Jones CJ, Edwards KJ, Castaglione S, Winfield MO, Sala F, Wiel C van de, et al. (1997). Reproducibility testing of RAPD, AFLP and SSR markers in plants by a network of European laboratories. Mol Breed 3: 381–390.

Jones M, Wagner R, Radman M (1987). Repair of a mismatch is influenced by the base composition of the surrounding nucleotide sequence. Genetics 115: 605–610.

Kambhampati S, Black WC, Rai KS, Sprenger D (1990). Temporal variation in genetic structure of a colonising species: *Aedes albopictus* in the United States. Heredity (Edinb) 64: 281–7.

Kambhampati S, Black WC, Rai KS (1991). Geographic origin of the US and Brazilian *Aedes albopictus* inferred from allozyme analysis. Heredity (Edinb) 67: 85–93.

Kambhampati S, Rai KS, Verleye DM (1992). Frequencies of mitochondrial DNA haplotypes in laboratory cage populations of the mosquito, *Aedes albopictus*. Genetics 132: 205–209.

Kamgang B, Brengues C, Fontenille D, Njiokou F, Simard F, Paupy C (2011). Genetic structure of the tiger mosquito, *Aedes albopictus*, in Cameroon (Central Africa). PLoS One 6: e20257.

Kamgang B, Ngoagouni C, Manirakiza A, Nakouné E, Paupy C, Kazanji M (2013). Temporal Patterns of Abundance of *Aedes aegypti* and *Aedes albopictus* (Diptera: Culicidae) and Mitochondrial DNA Analysis of *Ae. albopictus* in the Central African Republic. PLoS Negl Trop Dis 7: e2590

Kittayapong P, Baimai V, O'Neill SL (2002). Field prevalence of Wolbachia in the mosquito vector *Aedes albopictus*. Am J Trop Med Hyg 66: 108–111.

Kraemer MUG, Sinka ME, Duda KA, Mylne A, Shearer FM, Barker CM, et al. (2015). The global distribution of the arbovirus vectors *Aedes aegypti* and *Ae. albopictus*. Elife 4: e08347.

Kumar A, Rai KS (1990). Intraspecific variation in nuclear DNA content among world populations of a mosquito, *Aedes albopictus* (Skuse). Theor Appl Genet 79: 748–52.

Lacroix R, Delatte H, Hue T, Reiter P (2009). Dispersal and Survival of Male and Female *Aedes albopictus* (Diptera: Culicidae) on Réunion Island. J Med Entomol 46: 1117–1124.

Lenormand T, Bourguet D, Guillemaud T, Raymond M (1999). Tracking the evolution of insecticide resistance in the mosquito Culex pipiens. Nature 400: 861–4.

Liew C, Curtis CF (2004). Horizontal and vertical dispersal of dengue vector mosquitoes, *Aedes aegypti* and *Aedes albopictus*, in Singapore. Med Vet Entomol 18: 351–60.

Lourenço-de-Oliveira R, Vazeille M, Filippis AMB, Failloux AB (2003). Large genetic differentiation and low variation in vector competence for dengue and yellow fever viruses of *Aedes albopictus* from Brazil, the United States and the Cayman Islands. Am J Trop Med Hyg 60: 105–114.

Maia RT, Scarpassa VM, Maciel-Litaiff LH, Tadei WP (2009). Reduced levels of genetic variation in *Aedes albopictus* (Diptera: Culicidae) from Manaus, Amazonas State, Brazil, based on analysis of the mitochondrial DNA ND5 gene. Genet Mol Res 8: 998–1007.

Manni M, Gomulski LM, Aketarawong N, Tait G, Scolari F, Somboon P, et al. (2015). Molecular markers for analyses of intraspecific genetic diversity in the Asian Tiger mosquito, *Aedes albopictus*. Parasit Vectors 8: 188.

Marini F, Caputo B, Pombi M, Tarsitani G, della Torre A (2010). Study of *Aedes albopictus* dispersal in Rome, Italy, using sticky traps in mark-release-recapture experiments. Med Vet Entomol 24: 361–8.

McCoy KDD (2008). The population genetic structure of vectors and our understanding of disease epidemiology. Parasite 15: 444–448.

Medley KA, Jenkins DG, Hoffman EA (2015). Human-aided and natural dispersal drive gene flow across the range of an invasive mosquito. Mol Ecol 24: 284–295.

Medlock JM, Hansford KM, Schaffner F, Versteirt V, Hendrickx G, Zeller H, et al. (2012). A review of the invasive mosquitoes in Europe: ecology, public health risks, and control options. Vector Borne Zoonotic Dis 12: 435–447.

Minard G, Tran F-H, Tran-van V, Goubert C, Bellet C, Lambert G, et al. (2015). French invasive Asian tiger mosquito populations harbor reduced bacterial microbiota and genetic diversity compared to Vietnamese autochthonous relatives. Frontiers in Microbiology 6. doi:10.3389/fmicb.2015.00970.

Moore CG (1999). *Aedes albopictus* in the United-States: Current status and prospects for further spread. J Am Mosq Control Assoc 15: 221–227. Morales Vargas RE, Phumala-Morales N, Tsunoda T, Apiwathnasorn C, Dujardin JP (2013). The phenetic structure of *Aedes albopictus*. Infect Genet Evol 13: 242–251.

Mousson L, Dauga C, Garrigues T, Schaffner F, Vazeille M, Failloux A-B (2005). Phylogeography of Aedes (Stegomyia) aegypti (L.) and Aedes (Stegomyia) albopictus (Skuse) (Diptera: Culicidae) based on mitochondrial DNA variations. Genet Res 86: 1–11.

Mutebi J-P, Black WC, Bosio CF, Sweeney WP, Craig GB (1997). Linkage Map for the Asian Tiger Mosquito Aedes (Stegomyia) albopictus Based on SSCP Analysis of RAPD Markers. J Hered 88: 489–494.

O'Donnell D, Armbruster P (2010). Inbreeding depression affects life-history traits but not infection by Plasmodium gallinaceum in the Asian tiger mosquito, *Aedes albopictus*. Infect Genet Evol 10: 669–77.

Van Oosterhout C, Hutchinson WF, Wills DPM, Shipley P (2004). MICRO-CHECKER: Software for identifying and correcting genotyping errors in microsatellite data. Mol Ecol Notes 4: 535–538.

Van Oosterhout C, Weetman D, Hutchinson WF (2006). Estimation and adjustment of microsatellite null alleles in nonequilibrium populations. Mol Ecol Notes 6: 255–256.

Paupy C, Delatte H, Bagny L, Corbel V, Fontenille D (2009). *Aedes albopictus*, an arbovirus vector: from the darkness to the light. Microbes Infect 11: 1177–85.

Paupy C, Girod R, Salvan M, Rodhain F, Failloux AB (2001). Population structure of *Aedes albopictus* from La Réunion Island (Indian Ocean) with respect to susceptibility to a dengue virus. Heredity (Edinb) 87: 273–283.

Paupy C, Ollomo B, Kamgang B, Moutailler S, Rousset D, Demanou M, et al. (2010). Comparative Role of *Aedes albopictus* and *Aedes aegypti* in the Emergence of Dengue and Chikungunya in Central Africa. Vector-Borne Zoonotic Dis 10: 259–266.

Perez T, Albornoz J, Dominguez A (1998). An evaluation of RAPD fragment reproducibility and nature. Mol Ecol 7: 1347–1357.

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012). Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. PLoS One 7: e37135.

Poelchau MF, Reynolds J a, Elsik CG, Denlinger DL, Armbruster P a (2013). Deep sequencing reveals complex mechanisms of diapause preparation in the invasive mosquito, *Aedes albopictus*. Proc Biol Sci 280: 20130143.

Pongsiri A, Ponlawat A, Thaisomboonsuk B, Jarman RG, Scott TW, Lambrechts L (2014). Differential susceptibility of two field *Aedes aegypti* populations to a low infectious dose of dengue virus. PLoS One 9: e92971.

Porretta D, Gargani M, Bellini R, Calvitti M, Urbanelli S (2006). Isolation of microsatellite markers in the tiger mosquito *Aedes albopictus* (Skuse). Mol Ecol Notes 6: 880–881.

Porretta D, Mastrantonio V, Bellini R, Somboon P, Urbanelli S (2012). Glacial history of a modern invader: phylogeography and species distribution modelling of the Asian tiger mosquito *Aedes albopictus*. PLoS One 7: e44515.

Powell JR, Tabachnick WJ (2013). History of domestication and spread of *Aedes aegypti*–a review. Mem Inst Oswaldo Cruz 108 Suppl: 11–17.

Raharimalala FN, Ravaomanarivo LH, Ravelonandro P, Rafarasoa LS, Zouache K, Tran-Van V, et al. (2012). Biogeography of the two major arbovirus mosquito vectors, *Aedes aegypti* and *Aedes albopictus* (Diptera, Culicidae), in Madagascar. Parasit Vectors 5: 56.

Rao PN, Rai KS (1987). Inter and intraspecific variation in nuclear DNA content in Aedes mosquitoes. Heredity (Edinb) 59: 253–258.

Reiter P, Sprenger D (1987). The used tire trade: a mechanism for the worldwide dispersal of container breeding mosquitoes. J Am Mosq Control Assoc 3: 494–501.

Rezza G, Nicoletti L, Angelini R, Romi R, Finarelli AC, Panning M, et al. (2007). Infection with chikungunya virus in Italy: an outbreak in a temperate region. Lancet 370: 1840–6.

Roche B, Léger L, L'Ambert G, Lacour G, Foussadier R, Besnard G, et al. (2015). The Spread of *Aedes albopictus* in Metropolitan France: Contribution of Environmental Drivers and Human Activities and Predictions for a Near Future. PLoS One 10: e0125600.

Shaikevich E, Talbalaghi A (2013). Molecular Characterization of the Asian Tiger Mosquito *Aedes albopictus* (Skuse) (Diptera: Culicidae) in Northern Italy. ISRN Entomol 2013: 1–6.

Shaw PW, Turan C, Wright JM, O'Connell M, Carvalho GR (1999). Microsatellite DNA analysis of population structure in Atlantic herring (Clupea harengus), with direct comparison to allozyme and mtDNA RFLP analyses. Heredity (Edinb) 83: 490–499.

Sinkins SP, Braig HR, O'Neill SL (1995). Wolbachia superinfections and the expression of cytoplasmic incompatibility. Proc Biol Sci 261: 325–330.

Sprenger D, Wuithiranyagool T (1986). The discovery and distribution of *Aedes albopictus* in Harris County, Texas. J Am Mosq Control Assoc 2: 217–219.

Teixeira MG, Costa M da CN, Barreto F, Barreto ML (2009). Dengue: twenty-five years since reemergence in Brazil. Cad Saude Publica 25: S7–S18.

Urbanelli S, Bellini R, Carrieri M, Sallicandro P, Celli G (2000). Population structure of *Aedes albopictus* (Skuse): the mosquito which is colonizing Mediterranean countries. Heredity (Edinb) 84: 331–337.

Urbanski J, Mogi M, O'Donnell D, DeCotiis M, Toma T, Armbruster P (2012). Rapid adaptive evolution of photoperiodic response during invasion and range expansion across a climatic gradient. Am Nat 179: 490–500.

Usmani-Brown S, Cohnstaedt L, Munstermann LE (2009). Population Genetics of *Aedes albopictus* (Diptera: Culicidae) invading populations, using mitochondrial nicotinamide adenine dinucleotide dehydrogenase subunit 5 sequences. Ann Entomol Soc Am 102: 144–150.

Vazeille M, Mousson L, Rakatoarivony I, Villeret R, Rodhain F, Duchemin JB, et al. (2001). Population genetic structure and competence as a vector for dengue type 2 virus of *Aedes aegypti* and *Aedes albopictus* from Madagascar. Am J Trop Med Hyg 65: 491–497.

Wattier R, Engel CR, Saumitou-Laprade P, Valero M (1998). Short allele dominance as a source of heterozy-

gote deficiency at microsatellite loci: Experimental evidence at the dinucleotide locus Gv1CT in Gracilaria gracilis (Rhodophyta). Mol Ecol 7: 1569–1573.

Werren JH, Zhang W, Guo LR (1995). Evolution and phylogeny of Wolbachia: reproductive parasites of arthropods. Proc Biol Sci 261: 55–63.

Zawani MKN, Abu HA, Sazaly AB, Zary SY, Darlina MN (2014). Population genetic structure of *Aedes albopictus* in Penang, Malaysia. Genet Mol Res 13: 8184–96.

Zhong D, Lo E, Hu R, Metzger ME, Cummings R, Bonizzoni M, et al. (2013). Genetic analysis of invasive *Aedes albopictus* populations in Los Angeles County, California and its potential public health impact. PLoS One 8: e68586.

Zhou W, Rousset F, O'Neil S (1998). Phylogeny and PCR-based classification of Wolbachia strains using wsp gene sequences. Proc Biol Sci 265: 509–515.

Žitko T, Kovaćić A, Desdevises Y, Puizina J (2011). Genetic variation in East-Adriatic populations of the Asian tiger mosquito, *Aedes albopictus* (Diptera: Culicidae), inferred from NADH5 and COI sequence variability. Eur J Entomol 108: 501–508.

Zouache K, Fontaine A, Vega-Rua A, Mousson L, Thiberge J-M, Lourenco-De-Oliveira R, et al. (2014). Three-way interactions between mosquito population, viral strain and temperature underlying chikungunya virus transmission potential. Proc Biol Sci 281: 20141078.

Zouache K, Raharimalala FN, Raquin V, Tran-Van V, Raveloson LHR, Ravelonandro P, et al. (2011). Bacterial diversity of field-caught mosquitoes, *Aedes albopictus* and *Aedes aegypti*, from different geographic regions of Madagascar. FEMS Microbiol Ecol 75: 377–89.

Zug R, Hammerstein P (2012). Still a host of hosts for Wolbachia: Analysis of recent data suggests that 40% of terrestrial arthropod species are infected. PLoS One 7: 7–9.

# Chapitre 2

# Assemblage et analyse du répétome d'*Aedes albopictus*

*LM – Tu as essayé Trinity ?*
*CG – Je pense que je ne vais pas avoir le temps..."*

– Lyon, printemps 2014

# Avant propos

Les différents éléments en main pour aborder la structure des populations d'*Ae. albopictus*, nous devions également développer les marqueurs nous permettant de réaliser le scan génomique. Comme nous l'avons vu précédemment, nous explorions deux principales méthodes, l'une basée sur le RAD-seq, et l'autre sur le polymorphisme d'insertion des éléments transposables, ou Transposon Display (TD).

Si le TD ne requiert pas de connaissances avancées sur un génome, il est néanmoins nécessaire de connaître la séquence et l'abondance de la ou des familles d'ET qui seront utilisées. Ces informations permettent d'estimer le nombre de marqueurs qui pourront être génotypés, et dans notre cas, nous cherchions à obtenir plusieurs milliers de ces locus par individu. Une autre donnée importante est, comme nous l'avons évoqué en introduction, de connaître l'état de conservation des différentes copies d'une famille d'ET au sein des génomes. En effet, le TD repose sur l'utilisation d'une amorce spécifique pour amplifier le plus grand nombre de copies d'une famille d'ET donnée. Il est donc important de cibler les familles dont les copies sont les plus similaires, afin de maximiser le nombre de locus amplifiés par la même PCR.

Afin de mener la recherche de telles familles chez le moustique tigre, nous disposions de lectures non assemblées, issues du projet de séquençage d'une souche isolée à la Réunion et piloté par Patrick Mavingui et Claire Valiente Moro du laboratoire d'écologie microbienne à Lyon. Nous avons procédé dans un premier temps à une analyse de ces séquences à l'aide d'un pipeline dédié spécifiquement à ces fins : Repeat Explorer (Novák *et al.* 2010). Cet outil compare deux à deux les lectures non assemblées et les agrège ensuite en fonction de leur recouvrement et de leur divergence nucléotidique. Lorsqu'un échantillon représentant moins d'une fois la séquence complète du génome est utilisé, seules les lectures provenant de régions répétées sont suffisamment nombreuses pour être agrégées puis assemblées. Il est ainsi possible de connaître la séquence des ET les plus répétés, d'estimer leur abondance, qui est proportionnelle au nombre de lectures agrégées pour une famille donnée, et de mesurer relativement l'état de conservation des copies en prenant en compte la distance nucléotidique entre les lectures d'une famille.

Cette méthode a rapidement permis d'identifier des familles candidates, pour lesquelles nous avons pu développer des amorces de TD. Par ailleurs, cette analyse a révélé le contenu important de ce génome en éléments répétés, et il nous a semblé pertinent de valoriser ces travaux en produisant une description plus approfondie du « répétome » d'*Ae. albopictus*, des ressources qui pourraient être utiles aux projets de séquençages en cours, ces séquences atteignant près de 50% du génome de la souche étudiée.

En cherchant à approfondir nos analyses, nous nous sommes heurtés à certaines limitations de la méthode Repeat Explorer, en particulier concernant le temps de calcul, la qualité de l'assemblage et l'annotation automatique des ET identifiées. C'est ainsi que,

suivant l'idée soufflée par Laurent Modolo, nous avons essayé d'utiliser l'assembleur de données RNA-seq Trinity à la place de Repeat Explorer. Cet outil s'est révélé particulièrement habile pour identifier les ET présents au sein d'échantillons de lectures non assemblées, selon les même hypothèses que la méthode Repeat Explorer. Ceci faisant, les résultats d'assemblage se sont trouvés nettement améliorés, notamment concernant la détection des séquences alternatives (insertions, délétions) qui peuvent exister au sein d'une famille d'ET.

Nous avons ensuite construit un pipeline complet, automatisant les différentes étapes d'échantillonnage, d'assemblage des lectures, d'annotation et de quantification des familles d'ADN répétés, ainsi qu'une estimation de l'âge relatif des différentes familles d'ET, ce qui est inédit pour l'analyse de données non assemblées. Cette étape nous a notamment permis de reproduire les analyses et de diffuser la méthode. C'est ainsi qu'est né dnaPipeTE pour « de-novo assembly and annotation pipeline for transposable elements ».

L'article présenté dans ce chapitre décrit la méthode en détails, ainsi que l'analyse approfondie du répétome du moustique tigre, en comparaison avec le moustique de la fièvre jaune *Aedes aegypti*.

Enfin, nous proposons la première description exhaustive de la fraction répétée du génome du moustique tigre, qui atteint 50% du génome et dont les deux tiers, soit près de 30% du total, est composé d'ET. Nous avons aussi montré que les ET détectés chez *Ae. albopictus* pouvaient avoir été récemment, voire être toujours particulièrement actifs au sein du génome.

GBE

Genome Biology and Evolution

SMBE

# De Novo Assembly and Annotation of the Asian Tiger Mosquito (*Aedes albopictus*) Repeatome with dnaPipeTE from Raw Genomic Reads and Comparative Analysis with the Yellow Fever Mosquito (*Aedes aegypti*)

Clément Goubert[1,2,3], Laurent Modolo[1,2,3], Cristina Vieira[1,2,3], Claire Valiente Moro[2,3,4], Patrick Mavingui[2,3,4,5], and Matthieu Boulesteix[1,2,3,*]

[1]Laboratoire de Biométrie et Biologie Évolutive, UMR 5558, CNRS, INRIA, VetAgro Sup, Villeurbanne, France

[2]Université de Lyon 1, Villeurbanne, France

[3]Université de Lyon, Lyon, France

[4]Ecologie Microbienne, UMR 5557, CNRS, USC INRA 1364, VetAgro Sup, FR41 BioEnvironment and Health, Villeurbanne, France

[5]Université de La Réunion, UMR PIMIT, CNRS 9192, INSERM 1187, IRD 249

*Corresponding author: E-mail: matthieu.boulesteix@univ-lyon1.fr.

**Associate editor:** Josefa Gonzalez

**Accepted:** March 6, 2015

## Abstract

Repetitive DNA, including transposable elements (TEs), is found throughout eukaryotic genomes. Annotating and assembling the "repeatome" during genome-wide analysis often poses a challenge. To address this problem, we present dnaPipeTE—a new bioinformatics pipeline that uses a sample of raw genomic reads. It produces precise estimates of repeated DNA content and TE consensus sequences, as well as the relative ages of TE families. We shows that dnaPipeTE performs well using very low coverage sequencing in different genomes, losing accuracy only with old TE families. We applied this pipeline to the genome of the Asian tiger mosquito *Aedes albopictus*, an invasive species of human health interest, for which the genome size is estimated to be over 1 Gbp. Using dnaPipeTE, we showed that this species harbors a large (50% of the genome) and potentially active repeatome with an overall TE class and order composition similar to that of *Aedes aegypti*, the yellow fever mosquito. However, intraorder dynamics show clear distinctions between the two species, with differences at the TE family level. Our pipeline's ability to manage the repeatome annotation problem will make it helpful for new or ongoing assembly projects, and our results will benefit future genomic studies of *A. albopictus*.

**Key words:** transposable elements, repeated DNA, TE analysis, *Aedes albopictus*, Trinity, bioinformatic pipeline.

## Introduction

Repeated DNA, including transposable elements (TEs), is widespread within eukaryotic genomes. In such a "repeatome," the spread of TEs, which might bear coding sequences and can reach thousands of base pairs in length, contributes substantially to genomic size and evolution. Because of their ability to insert within genes or regulatory regions and to cause ectopic recombination due to their repetitive nature, TEs are assumed to be frequently deleterious to their hosts (Goodier and Kazazian 2008; Beck et al. 2011; Vela et al. 2014). However, an increasing number of studies have shown that TE insertions can sometimes be adaptive and can be co-opted by their host genomes (Rebollo et al. 2010; Casacuberta and González 2013). Thus, understanding genomic evolution demands a comprehensive knowledge of TE composition within the genome, as well as of their dynamics and interactions with host genome. To this end, genome annotations that include TE annotation and quantification are crucial.

In the current era of short-read sequencing, the assembly of genomes bearing a significant amount of repeated sequence is a complex task. Reads overlapping a repeated element might correspond to several positions in the genome

and thus can be misplaced and can produce chimeric assembly. Therefore, repeats produce a large number of short contigs that cannot be properly positioned or annotated within the assembly. Accordingly, the quality of the assembly for TEs is often poor and can result in underrepresented and/or incorrect annotation of their sequences (Modolo and Lerat 2014).

The Asian tiger mosquito *Aedes albopictus* (Diptera: Culicidae) presents a striking example of a genome that is difficult to assemble due to its repeatome. This species—a vector of Dengue and Chikungunya viruses that is often viewed as one of the most threatening invasive species in the world—still has not had its genome sequence released, even though several projects have been aimed at this task over the last few years (see Bonizzoni et al. 2013 for a review). *Aedes aegypti*, the closest species whose genome has been fully sequenced and annotated, possesses a similar genome size, and repeated DNA comprises more than 50% of its genome. Unlike *A. albopictus*, the whole genome of *A. aegypti* has been fully sequenced using Sanger technology, which produces longer reads than current Next-Generation Sequencing (NGS) methods and therefore allowed the construction of a large library of TEs and repeats (Nene et al. 2007). Moreover, intraspecies variation of the *A. albopictus* genome size—ranging from 0.62 to 1.66 pg—has been suggested (Rao and Rai 1987; Kumar and Rai 1990), supporting the hypothesis of a significant amount of TE activity, with more copies present in some populations than in others (McLain et al. 1987; Black et al. 1988). However, no study is currently aimed at finding and quantifying TEs in a comprehensive manner in this species.

Several bioinformatic solutions now enable the de novo assembly of TE sequences directly from NGS genomic data sets without the need for a reference genome. These methods assume that reads belonging to TEs or other repetitive DNAs are overrepresented among the sequenced reads. Current pipelines such as RepARK (Koch et al. 2014) and TEdna (Zytnicki et al. 2014) use whole NGS genomic data sets or only the unassembled reads left after a genome assembly. These two programs use overrepresented k-mers to assemble TE sequences: Velvet (Zerbino and Birney 2008) or CLC (CLCbio, http://www.clcbio.com/products/clc-assembly-cell/, last accessed April 13, 2015) are used in RepARK, and an implementation of a de Bruijn graph assembler is used in TEdna. Although these programs are dedicated to TE assembly, they do not allow repeat quantification or annotation. An alternative way to explore a genome's repetitive content is to use low coverage sequencing. In such data sets, only TEs and other repetitive DNA sequences are expected to have a sufficient representation in the pool of reads to be assembled. For example, in average, for a sample with 0.1× coverage, only sequences that are present at least 10 times within the genome can be assembled. Based on this principle, the RepeatExplorer (RE) pipeline (Novák et al. 2010) was designed to cluster and then assemble similar reads from a small uniform genomic sample in order to retrieve repeats. In a uniform genomic sample, the proportion of reads assigned to a given cluster directly corresponds to the proportion of reads assigned to the relevant TE family in the genome. In addition to computing a direct quantification of each repeat family, RE can annotate repeat families using RepeatMasker (RM) and protein domain search (Smit AFA, Hubley R, Green P. RepeatMasker Open-3.0. 1996–2010, http://www.repeatmasker.org, last accessed April 13, 2015). However, although the RE pipeline can process NGS data sets, most of the tools it uses are not designed for this type of data, especially during the assembly step performed by CAP3 (Huang 1999)—a Bacterial Artificial Chromosome (BAC)-clone sequence type assembler.

Here, we present a new pipeline, dnaPipeTE (De Novo Assembly and Annotation Pipeline for Transposable Elements), that combines previous methods by allowing fast and accurate assembly of repeat sequences from a small genomic sample with dedicated NGS tools and by performing quantification and annotation of TEs and repeats for comparative analysis. The cornerstone of dnaPipeTE is the use of Trinity (Grabherr et al. 2011)—originally designed for RNAseq data assembly—to assemble repeats from low-coverage genomic data sets, which produce complete repeat sequences and enable the recovery of alternative consensuses within one TE family. Our pipeline also performs an automatic annotation of repeats using RM and the Repbase database (Jurka et al. 2005) and produces different data and figures for the quantification of repeats. We also implemented a computation of the TE age distribution for the most recent copies, using the divergence between reads and contigs.

With this pipeline and annotations from known TEs, we aimed to 1) estimate the number of repeated DNAs in *A. albopictus*, 2) annotate and quantify the diversity of TEs in its genome, and 3) compare this repeatome with that of *A. aegypti*, to infer the dynamics of TEs since the divergence of these two species.

## Materials and Methods

### dnaPipeTE: A Pipeline to Assemble, Annotate, and Quantify Repetitive Sequences from Small Unassembled NGS Data Sets

dnaPipeTE is a fully automated pipeline designed to assemble and quantify repeats from genomic NGS reads. It is freely available for download at https://lbbe.univ-lyon1.fr/-dnaPipeTE-.html (under the GPLv3). Figure 1 shows the main steps in the dnaPipeTE pipeline. Our pipeline takes as input a FASTQ (Cock et al. 2010) file containing quality filtered short reads. dnaPipeTE then performs uniform samplings of the reads to produce low coverage data sets used during analysis. The samples must represent less than 1× coverage to avoid the assembly of nonrepeated genome content; using

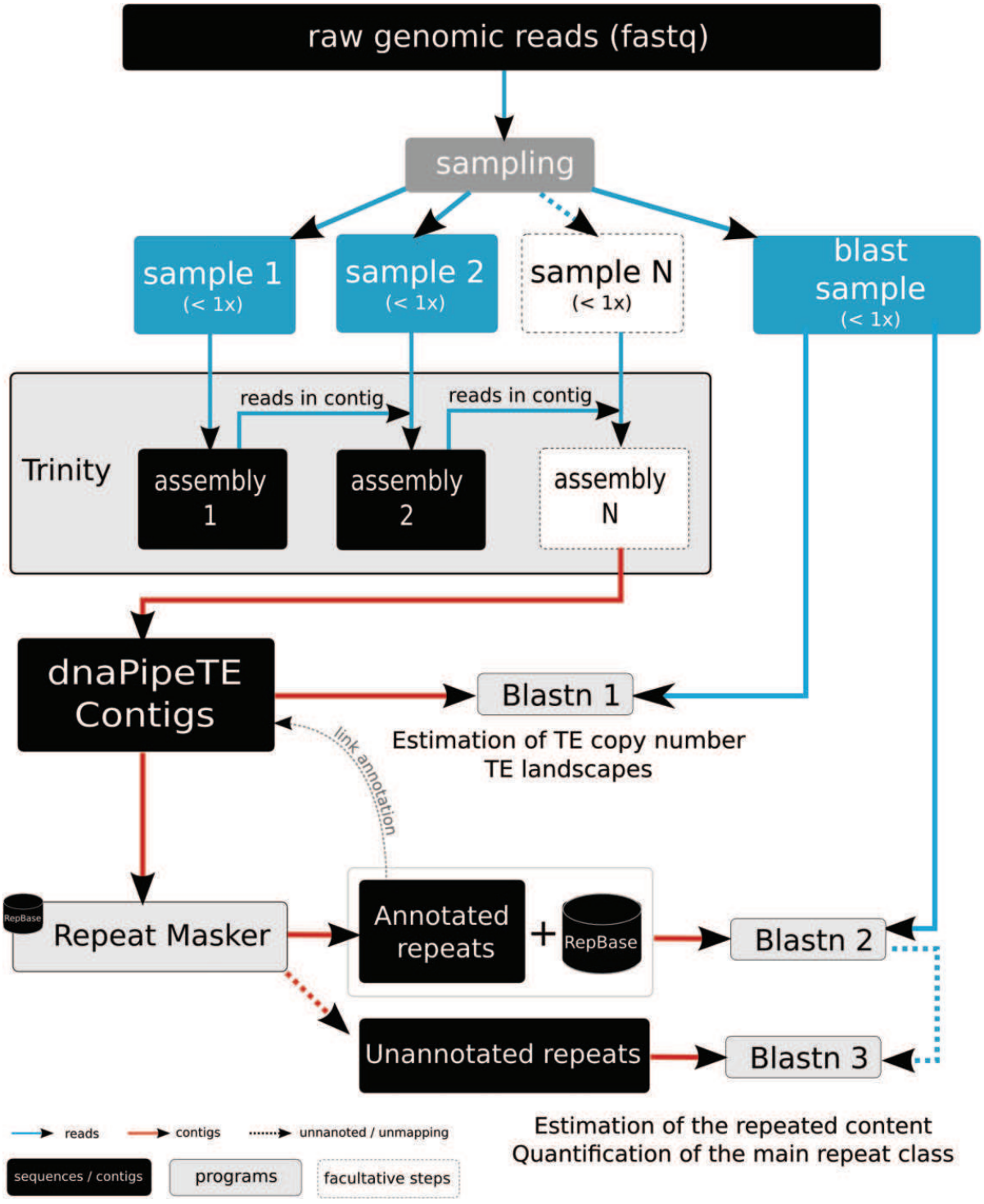



**Fig. 1.**—Overview of the dnaPipeTE pipeline. First, genomic reads in FASTQ format are sampled. Then, assembly of repeats is performed using two or more iterations of Trinity. For each iteration, the previously assembled reads are added to the next sample to improve the repeat assembly. In the next step, assembled contigs are annotated using RepeatMasker. Finally, reads from the "BLAST sample" are blasted against all the contigs to estimate the relative abundance of each assembled repeat and to compute the TE landscape. In a second BLAST, the same sample is successively blasted against the annotated contigs joined to the Repbase library, then with the unannotated contigs in order to retrieve copies that would not have been assembled and to obtain a more global repeat content estimation. See text for additional details.

a sample size of less than 0.25× of the genome is often sufficient to obtain a precise estimate of the repeated content (see supplementary fig. S1, Supplementary Material online, for examples with 0.1× and 0.25×). dnaPipeTE requires at least three samples of the original genomic data set: Two for the assembly step and an independent third used for the quantification steps. Our pipeline is currently designed to use only single-end reads because training analyses showed that using paired-end reads could produce chimeras during repeat assembly (data not shown). We developed dnaPipeTE using 100-bp reads, which are currently the most frequently generated NGS data sets, but our implementation would work with any read size.

### *Repeat Assembly with Trinity*

After uniform sampling of the reads, dnaPipeTE builds contigs from the repeated sequences using Trinity. In an RNAseq experiment, a given gene can produce different transcripts, and the Trinity software is equipped to handle alternative transcripts with a hierarchical procedure: after identifying a "gene" (a subpart of the assembly graph), Trinity can produce different contigs that represent all the alternative transcripts of this gene. Similarly, TE copies from the same family, which may display an accumulation of mutations, deletions, insertions, or other structural changes, are treated by Trinity as alternative sequences of the same gene (TE family). Thus, with Trinity one can recover complete alternative consensus sequences from a given TE family. Retrieving good consensus increases the ability to perform an accurate estimation of TE abundance by improving read mapping to TEs. The rarest elements in the genome are predicted to generate few (or no) reads in the subset samples; thus, dnaPipeTE performs iterative runs of Trinity using new samples to decrease such risk. The first run uses a first sample; then, any reads mapping to k-mer contigs belonging to repeats ("inchworm" contigs; see Trinity manual) are added to a second independent sample, and Trinity is performed one more time. Each iteration enriches the number of reads associated with a repeat in the next sample and allows the recovery of more and larger contigs (some examples are given in supplementary Material, Supplementary Material online). In the case of *A. albopictus* sequences, our tuning experiments showed that two iterations performed on a data set with 0.1× coverage ensured the best assembly N50 and that supplementary iteration showed no significant improvement in the quality of the assembly (supplementary fig. S2, Supplementary Material online). In the latest versions of Trinity (≥r20140717), contigs are built from "clusters" that correspond to units of the de Bruijn graph made during the assembly. These clusters are divided into genes and finally "isoforms" that represent the alternative transcripts of a gene in RNAseq studies. Applied to low-coverage DNA data, one gene ideally represents one repeat family, in which isoforms are structural variant copies belonging to one family (copies with insertions or deletions for example) or to closely related families. An isoform present in Trinity.fasta output following all iterations of the Trinity program is referred to as a "dnaPipeTE contig." During the assembly step in dnaPipeTE, Trinity (version r20140717) was used with default parameters for single-end reads, with the exception of the minimum coverage to join k-mer contigs set to 1 to retain contigs from low copy repeats (Haas B, personal communication).

### *Contig Annotation with RepeatMasker*

After the assembly step, dnaPipeTE contigs are annotated using RM, for which a built-in or custom repeat library can be specified. Following the 80-80-80 rule proposed by Wicker et al. (2007), contigs with 80% query coverage on 80% of subjects (databases) were stored as "full-length," and queries with 80% hits on fewer than 80% of subjects were stored as "partial" (fig. 2). Of the other contigs annotated by RM, only the order information (according to Wicker et al. 2007 classification)— Long Terminal Repeat (LTR), Long INterspersed Element (LINE), Short INterspersed Element (SINE), DNA, Miniature Inverted-repeat Transposable Elements (MITEs) (short TEs harboring terminal inverted repeats but without coding sequences), Ribosomal RNA, low complexity, and simple/tandem repeats—is retained. For our analysis, we used the Repbase libraries (version 2014-01-31 downloaded from http://www.girinst.org/, last accessed April 13, 2015) and the TEFam library (accessed at http://tefam.biochem.vt.edu/tefam/index.php, last accessed April 13, 2015). RM (version open-4.0.5) parameters were set to default values, slow-research mode with the NCBI BLAST program (RMBLASTN program, NCBI BLAST 2,2,23+), and only the best hit was kept following dnaPipeTE contig analysis, as determined by the highest Smith–Waterman score provided by RM.

### *Repeat Quantification*

For quantifying the repeats, BLASTN software (Altschul et al. 1990) was found to perform better than classic short-read aligners such as Bowtie2 (Langmead and Salzberg 2012). Indeed, the divergence between a dnaPipeTE contig—that is, a consensus sequence for a repeat family—and its reads belonging to different copies can be higher than the divergence between a gene or a transcript and its reads, and requires a more sensitive approach. During the "BLAST 1" step (fig. 1), reads from the "BLAST" sample are matched against all the dnaPipeTE contigs to estimate the genome proportion of each assembled repeat. However, we cannot quantify the unassembled repeats during this step. Thus, to obtain an overall estimation of repeat content, the BLAST sample is first matched against a database composed of the annotated contigs of dnaPipeTE and the repeat library in order to recover reads associated with misassembled or missing repeats

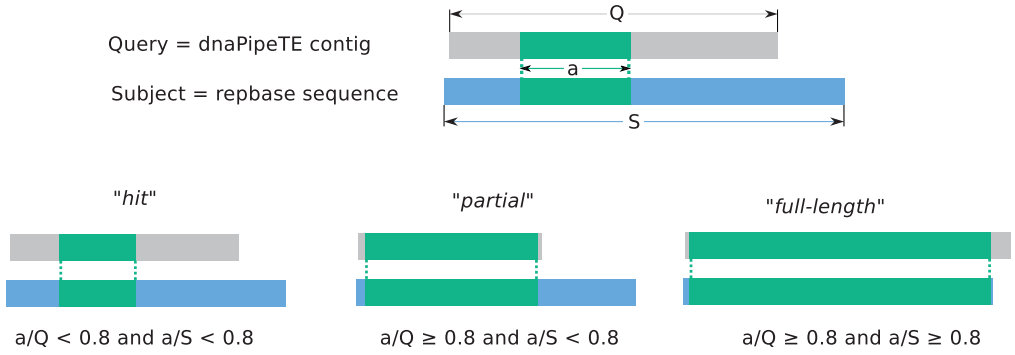



**Fig. 2.**—Classification procedure of RepeatMasker annotation for the dnaPipeTE contigs. According to the alignment overlap between the query (a/Q) and the subject (a/S), the dnaPipeTE contigs are annotated as one of the three categories. "Hit" is the weakest annotation, while partial and full-length indicate that the dnaPipeTE contig has annotated along more than 80% of its length.

("BLASTN 2," fig. 1). Then, the unmapped reads are matched against the unannotated contigs supplied by dnapipeTE ("BLASTN 3," fig. 1), and the remaining reads are assumed to belong to nonrepeated sequences. We use the BLAST sample for both estimations, and reads are mapped using discontinuous BLASTN (NCBI BLAST 2.2.29+), which keeps matches with 80% minimum identity and only the best hit per read. To speed-up computation, dnaPipeTE uses GNU Parallel (version 20 140 622) (Tange 2011) to parallelize BLASTN runs.

Finally, the divergence computed between one read and its contigs during the BLAST 1 step is used as a proxy of the divergence time between TE copies in a given family. This proxy is shown to be relevant compared with previous analyses of TE age distribution that used Kimura distances from a full-length TE copy and its consensus sequence in Repbase ("TE Landscapes," http://www.repeatmasker.org/ (last accessed April 13, 2015); several examples are given in supplementary fig. S3, Supplementary Material online).

### *Efficiency of dnaPipeTE*

Prior to *A. albopictus* genome analysis, we tested the efficiency of dnaPipeTE on well-annotated genomes that varied in size and TE content. We used available Illumina reads from the species *Drosophila melanogaster* (Diptera: Drosophilidae), *Anopheles gambiae* (Diptera: Culicidae), *Caenorhabditis elegans* (Rhabditida: Rhabditidae), *Ciona intestinalis* (Enterogona: Cionidae), *Gasterosteus aculeatus* (Gasterosteiformes: Gasterosteidae), and *A. aegypti*—the closest fully sequenced species to *A. albopictus*. We also tested the behavior of dnaPipeTE on older repeatomes, such as that of the human genome (*Homo sapiens*), in which copies of one TE family are highly divergent. All data management information and references are given in supplementary table S1, Supplementary Material online.

## Analysis of the *A. albopictus* Repeatome and Comparison with *A. aegypti*

### *Genomic Data*

The two mosquito genomes were sequenced with Illumina NGS technology (Illumina HiSeq2000). The *A. albopictus* strain originated from La Reunion Island, Indian Ocean. Genomic DNA was prepared from four female individuals of generation F5 bred in an insectarium. Sequencing generated 440.2 million 100-bp paired-end reads (ProfilXpert platform, Lyon, France). A total sample of 4,243,902 single-end reads was also generated (R1's were used). *Aedes aegypti* female genomic reads (SRR871496; strain Liverpool; 213.4 million 100-bp paired-end reads; ~16.4× coverage, Virginia Tech) were downloaded from the short-read archive collection (http://www.ncbi.nlm.nih.gov/sra, last accessed April 13, 2015); only the first read of each pair was used for analysis.

### *Read Preprocessing*

According to quality statistics, all reads were trimmed to 82 bp, keeping the nucleotides 10 through 91 in both *A. albopictus* and *A. aegypti* species. Then, sequences were filtered using FASTX-toolkit (http://hannonlab.cshl.edu/fastx_toolkit/, last accessed April 13, 2015) with a minimum 20 average Phred score on 90% of the reads. Finally, reads from mitochondrial DNA were removed from the data with Bowtie 2 software (version 2.1.0) under default parameters to map reads to the whole mitochondrial genome sequence for each *Aedes* species available through the NCBI website (http://www.ncbi.nlm.nih.gov/, last accessed April 13, 2015).

### *Aedes albopictus and A. aegypti Sampling*

In the literature, the genome size of *A. albopictus* is reported to be variable, ranging from 0.6 to 1.6 Gbp. Flow cytometry performed on the heads of *A. albopictus* females estimated the genome size of our sequenced strain to be 1.16 Gbp

(1.19 pg, unpublished data). The number of reads comprising the three independent samples used by dnaPipeTE was set to represent 0.1× of each genome. The subset sample of 4,243,902 reads (0.3×) was used to assemble TEs and repeats for *A. albopictus*, consisting of 2 samples of 0.1× genomic coverage for assembly and a third sample of 0.1× for the quantification step. This sample size was chosen after a preliminary analysis showed that 0.1× per Trinity run maximizes the assembly N50 for this genome (supplementary fig. S2, Supplementary Material online). We suggest that this will balance finding as many repeats as possible with limiting the assembly of nonrepeated DNA (noise). For *A. aegypti*, coverage was also set to 0.1×, using reads taken from the full sequencing experiment based on a genome size of 1.3 Gbp, according to the whole-genome assembly size and mean genome size estimations (Nene et al. 2007; Gregory, T.R. (2015); Animal Genome Size Database. http://www.genomesize.com, last accessed April 13, 2015).

### *TE Family Recovery and Quantification*

To cluster dnaPipeTE contigs into TE families, we used the cd-hit-est program from the CD-HIT suite (version 4.6.1) (Li and Godzik 2006) with local alignment and the greedy algorithm. We set the clustering parameters to group pairs of sequences with at least 80% of the shortest sequence aligned, with a minimum of 80% identity in the longest sequence (parameters -aS 0,8 -c 0,8 -G 0 -g 1). This method results in better performance than grouping contigs per Trinity gene or by RM annotation. In the first case, contigs from one Trinity gene could be joined when they shared a conserved fragment (such as a protein domain), even if they did not actually belong to the same TE family. In the second case, RM annotations include only the closest sequences known, and one sequence could easily match to multiple TE families. This method allowed us to report the most abundant repeats (in relative genome proportion) and to estimate the number of TE copies for fully assembled repeats (dnaPipeTE contigs full-length, see above).

We then estimated the copy number of the fully assembled repeats (table 1) using the following formula:

$$(n/N) \times (G/L)$$

where $n$ is the number of read-matching contigs from a TE family (contigs from one CD-HIT cluster), $N$ is the total number of reads in the BLAST sample, $G$ is the genome size in bp, and $L$ is the length of the representative sequence of the TE family (reference sequence of the CD-HIT cluster) in bp.

### *TE Transcriptional Activity*

To identify transcriptionally active TEs among the discovered repeats in *A. albopictus*, we mapped the *A. albopictus* transcriptome assembly (adult, embryo, and oocyte transcriptome merged reference assembly downloaded from http://www.albopictusexpression.org/, last accessed April 13, 2015) onto the dnaPipeTE contigs using BLAT. We filtered the results of the BLAT analysis such that only TE consensus sequences matching 80% of a transcriptome contig (minimum alignment 80 bp) with 80% minimum identity were retained.

### *Comparison between A. albopictus and A. aegypti*

To avoid annotation bias due to the abundance of reference sequences from *A. aegypti* in Repbase, we performed a second analysis with dnaPipeTE on *A. albopictus* and *A. aegypti* using a TE library devoid of reference sequences from *A. aegypti*. Then, we used BLAT to match cd-hit-clustered dnaPipeTE contigs between species in order to identify shared TE families. We filtered the results of the BLAT analysis such that alignments with at least 80 bp and 75% identity and only one reference contig per species were retained. Finally, for each species we summed the total number of reads in the cluster for which the references belonged. Thus, we obtained pairs of counts for putatively shared TE families.

### dnaPipeTE Comparison with RepeatExplorer

Compared with dnaPipeTE, RE requires only one sample for assembly and annotation. We thus ran it using the "BLAT" sample generated by dnaPipeTE for the *A. albopictus* data set, on which an estimation of repeated content and a quantification of the main repeat families is performed. Computations were performed online with the "clustering" tool of the RE Galaxy server (http://repeatexplorer.umbr.cas.cz/, last accessed April 13, 2015) with the following parameters: 44 bp (55% of the read length) minimum overlap for clustering, 0.01% cluster threshold for detailed analysis, 40 bp minimal overlap for read assembly and RepeatMasking against the "all" database. Computation time, contig number, N50, proportion of repeats in the sample, and percentage of annotation of the repeated content were calculated for comparison.

## Results

### Efficiency of dnaPipeTE

We report here the results obtained for *D. melanogaster* (fig. 3). Details and results from other species are presented in supplementary figures S1 and S3, Supplementary Material online. In *D. melanogaster*, as well as the other fully annotated genome tested, dnaPipeTE estimations for the different families of TEs are accurate when only a small subset sample of NGS sequencing reads was used as input (three samples of 0.25× coverage). The relative proportion of each TE order is respected in dnaPipeTE estimations. In *D. melanogaster*, however, the whole repeat content is underestimated (17.78% vs. 28.21%). For this species, our results indicate that dnaPipeTE seems to have underestimated the simple and tandem repeat content of the genome. For *A. aegypti* (supplementary fig. S1,

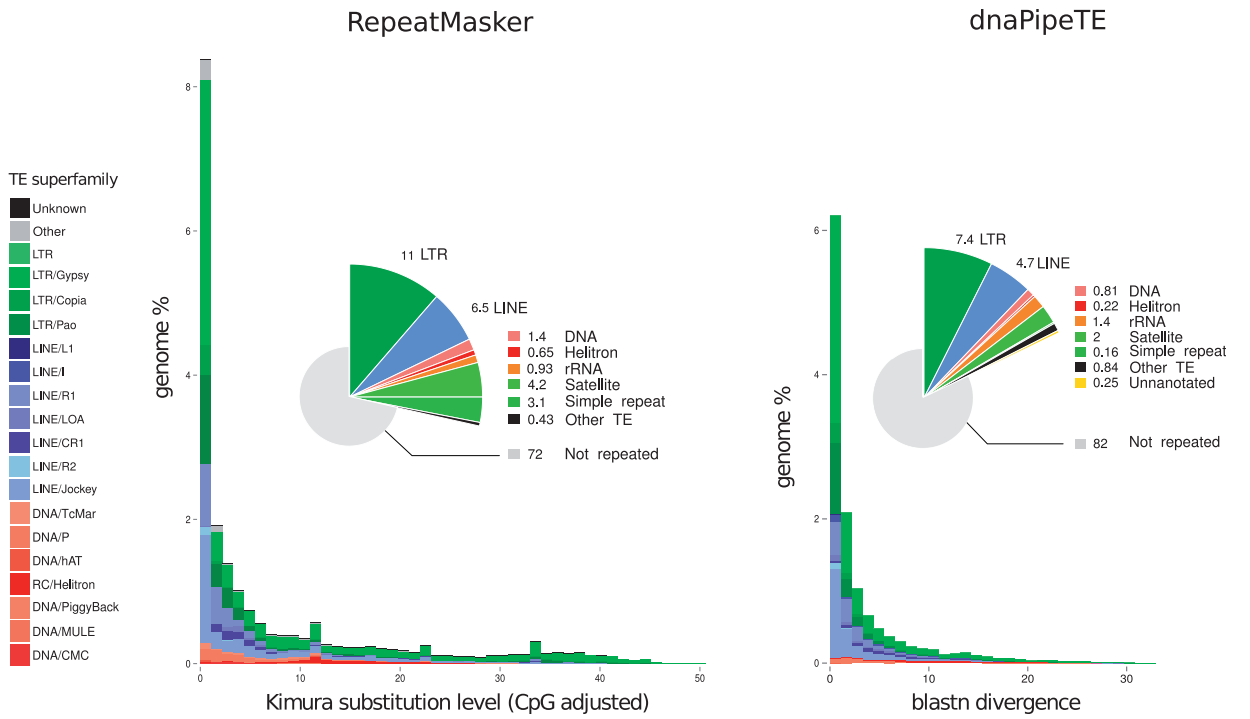



**Fig. 3.**—Relative genome proportions of the main repeat classes (pie charts) and TE landscapes (bar plots) from RepeatMasker on assembled genome (left) and dnaPipeTE (right, BLASTN with 0.25× genome coverage) for *Drosophila melanogaster* strain *w1118*. RepeatMasker analysis data were downloaded from http://repeatmasker.org and retranscribed according to the name used for annotation in dnaPipeTE.

Supplementary Material online), we estimate the TE content to be 45.6%, which is very close to the estimation of 47% made by Nene et al. from the assembled genome. Using genomes variable in size and TE content as benchmark, we also noticed that the more the genome is filled with repeated DNA, the less the number of Trinity iteration is needed, as well as the coverage provided as input.

Comparisons of TE age distributions obtained with dnaPipeTE (fig. 3 and supplementary fig. S3, Supplementary Material online) and those made from fully assembled genomes available on the RM website (TE landscapes) (http://repeatmasker.org/genomicDatasets/RMGenomicDatasets.html, last accessed April 13, 2015) were performed. These comparisons showed that dnaPipeTE provides a good estimate of the recent TE age distribution. As with other de novo TE assemblers, dnaPipeTE is limited in its ability to detect old TE families with degraded and divergent copies. For example, in *D. melanogaster* or *H. sapiens,* TEs with more than 30% divergence between reads and the consensus sequence are not identified (fig. 3 and supplementary fig. S4, Supplementary Material online). Our tuning tests show that dnaPipeTE performs well in the estimation of TE proportion and dynamics, with consensus-read divergence ranging from 0% to 15%, which is sufficient to compare closely related species and is close to the definition of a TE family as per the 80-80-80 rule (Wicker et al. 2007).

## *Aedes albopictus* Repeatome Analysis

### *Repeat Assembly with dnaPipeTE*

Assembly of the repeats produced 8,102 contigs with an N50 of 677 bp. Although no reference genome for *A. albopictus* exists at this point in time, dnaPipeTE was able to annotate 5,141 contigs including 949 "partial TEs" and 30 full-length elements. Among these, some full-length annotated dnaPipeTE contigs were found to represent different variants of the same family, including some internal deletions. Taking this into account, a total of 24 annotated families with full-length consensus sequences were quoted for *A. albopictus*.

### *Repeated DNA Content of A. albopictus*

dnaPipeTE reported that the repeatome of *A. albopictus* comprises 49.73% of the genome. Annotation of this repeated

**Table 1**

The Most Abundant Identified Repeat Families in *Aedes albopictus*

| Genome% | RM Annotation | RM Superfamily | dnaPipeTE Contig Size | Estimated Copy Number |
|---|---|---|---|---|
| 1.26% | Lian-Aa1 | LINE/LOA | 4,080 | 3586 |
| 1.25% | RTE Ele4 | LINE/RTE-BovB | 3,447 | 4203 |
| 1.16% | JAM1 | LINE/RTE-BovB | 2,356 | 5728 |
| 1.10% | R1_Ele1 | LINE/R1 | 5,797 | 2195 |
| 0.54% | RTE_Ele3 | LINE/RTE-BovB | 3,283 | 1911 |
| 0.41% | CACTA-3_AA | DNA/CMC-EnSpm | 1,626 | |
| 0.37% | TF001239_mTA_Ele24Aedes | MITE | 638 | |
| 0.33% | Chapaev3-2_AA | DNA/CMC-Chapaev-3 | 1,611 | |
| 0.29% | Loner_Ele2 | LINE/I | 6,335 | 526 |
| 0.28% | TF001239_mTA_Ele24_Aedes | MITE | 469 | |
| 0.28% | Loner Ele1 | LINE/I | 6,329 | 513 |
| 0.23% | Lian-Aa1 | LINE/LOA | 934 | |
| 0.23% | FEILAI_AA | S1NE/tRNA | 324 | 8215 |
| 0.22% | TF001248_mTA_E1e33_Aedes | MITE | 2,407 | 1071 |
| 0.18% | MSAT-1_AAe Satellite | | 2,133 | |
| 0.17% | RTE Ele5 | LINE/RTE-BovB | 2,642 | |
| 0.17% | Lian-Aa1 | LINE/LOA | 1,865 | 1053 |
| 0.17% | LSU-rRNADme | rRNA | 4,681 | |
| 0.16% | R1_Ele1 | LINE/R1 | 3,362 | |
| 0.16% | JAM1B_AAe | LINE/RTE-BovB | 793 | |
| 0.16% | LOA_Ele5 | LINE/LOA | 3,724 | 500 |
| 0.16% | TF001244_mTA_Ele29Aedes | MITE | 578 | |
| 0.15% | MSAT-2_AAe | Satellite | 1,301 | |
| 0.15% | TF001312_m8bp_Ele20_Aedes | MITE | 1,532 | |
| 0.15% | TF000681_m4bp_Ele5_Aedes | MITE | 674 | 2548 |
| 0.14% | CR1-50_AAe | LINE/CR1 | 678 | |
| 0.14% | Sola2-4_AAe | DNA/Sola | 1,232 | |
| 0.14% | TF001310_m8bp_E1e19_Aedes | MITE | 1,840 | |
| 0.14% | TF001280_otherMITEs_Ele7Aedes | MITE | 252 | |
| 0.13% | JAM1B_AAe | LINE/RTE-BovB | 424 | |
| 0.13% | MSAT-1_AAe | Satellite | 663 | |
| 0.13% | MSAT-2_AAe | Satellite | 575 | |
| 0.13% | Gecko | SINE/tRNA-I | 249 | 5967 |
| 0.13% | TF001295_mTA_Ele38c_Aedes | MITE | 1,377 | |
| 0.12% | MSAT-1AAe | Satellite | 204 | |
| 0.12% | TF001257_m4bp_E1e16_Aedes | MITE | 887 | |
| 0.12% | TF001280_otherMITEs_Ele7Aedes | MITE | 1,379 | |
| 0.12% | TF001313_otherMITEs_Ele27Aedes | MITE | 2,209 | |
| 0.12% | MSAT-1_AAe | Satellite | 852 | |
| 0.12% | TF000746_mTA_Ele22_Aedes | MITE | 557 | 2439 |
| 0.11% | LOA_Ele2B_AAe | LINE/LOA | 2,484 | |
| 0.11% | Sola1-3_AA | DNA/Sola | 349 | |
| 0.11% | otherMITEs_Ele11 | DNA/hAT-hATm | 421 | |
| 0.11% | TF001251_m3bp_Ele8a_Aedes | MITE | 900 | |

Note.—An estimation of copy number was made only for TEs identified as full-length elements and was based on the size of the dnaPipeTE reference contig after TE family clustering. RM annotation, repeat family hit found by RepeatMasker; RM superfamily, repeat superfamily name in Repbase.

DNA showed that TEs occupy 33.58% of the genome. Tandem repeats (satellites and microsatellites) occupy 8% (fig. 4), while unannotated repeats represent 7.23%. The most abundant repeats were Class II (DNA) transposons and LINE (Class I non-LTR) retrotransposons, followed by LTR retrotransposons and SINEs. Details regarding the most abundant repeat families are reported in table 1. The most abundant TE family in terms of genome percentage is a "Lian-like" LINE element (similar to *Lian-a1* in *A. aegypti*), which occupies 1.267% of the genome with 3,586 estimated copies (table 1). The most highly represented families in terms of copy number among the full-length elements annotated by dnaPipeTE are two LINE elements from the "Loner" superfamily, with more than 6,000 estimated

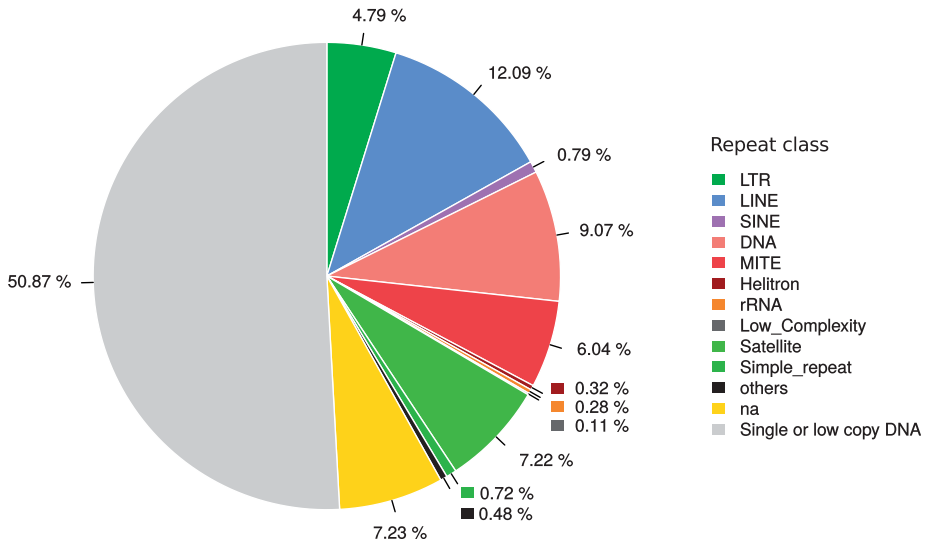



**Fig. 4.**—Relative genome proportions of the main repeat classes found in *Aedes albopictus* using dnaPipeTE, from a nucleotide BLAST of 1,414,634 reads (0.1×) against the repeat assemblies performed with a total of 2,829,268 reads (0.2×).

copies each. Thirteen other LINE families represent more than 0.10% of the genome each. Fourteen MITEs (non-autonomous Class II) also appear among the most repeated TE families.

In addition, we found using BLAT that 7,005 of the 8,102 dnaPipeTE contigs have significant hits with a sequence from the *A. albopictus* transcriptome assembly reported for adult, embryo, and oocyte (Poelchau 2011; http://www.albopictusexpression.org/ [last accessed April 13, 2015]; supplementary table S2, Supplementary Material online).

## Comparison of TE Dynamics between *A. albopictus* and *A. aegypti*

*Aedes albopictus* TE age distribution was compared with that of the yellow fever mosquito, *A. aegypti* (the only available assembled genome for the *Aedes* genus). We showed that in both species, most of the reads are highly similar to their respective dnaPipeTE contigs (fig. 5). This indicates that most of the detected TE families are recent and possess a high degree of similarity between their copies. This similarity is particularly strong for the detected LTR retrotransposons and, to a lesser extent, for the LINEs that are the most represented TEs in these distributions. Class II DNA transposons are less represented than expected in these comparisons, as their detection suffered from the removal of *A. aegypti* reference sequences from the library for comparison (fig. 4 for the full analysis in *A. albopictus* vs. fig. 5 for the interspecies comparison). Between species, the most striking result is that the genomic proportion of LINE/Jockey reads in *A. aegypti* is high and is composed of mostly recent but also some older TEs, while this family is much less abundant in *A. albopictus*, with less divergence between reads and contigs. In addition, the distribution of the read divergence of LINE/R1 elements is strongly concentrated at the left of the graphic (representing recent TE copies) in *A. aegypti*, while in *A. albopictus* the proportion of reads in superfamilies of higher divergence decreases more slowly (representing older TE copies).

The weak positive correlation between *A. aegypti* and *A. albopictus* in the genomic abundance of the shared families (fig. 6, $r^2 = 0.186$, $P < 0.01$ on the $\log_{10}$ scale) is mostly due to the less abundant families (<0.1% of the genome). Some families display very high differences, such as the *Juan*-A (LINE/Jockey retrotransposon) family which represents almost 3% of the genome proportion in *A. aegypti* but only 0.08% in *A. albopictus*, or *Copia_Ele122* which displays a 5-fold change between the two species, while *R1-Ele*1 and *RTE*-3 are good examples of the mirror case. Globally, very few shared families have the same genomic proportion, with the exception of *CACTA*-3 (DNA transposon) and, less markedly, *Jam*-1 or *Lian-Aa*1 (LINEs), which contrast the general trend.

## Comparison between dnaPipeTE and RepeatExplorer

Our pipeline dnaPipeTE operates on the same principles as RE to estimate, assemble, and annotate the repeatome of a species from a sample of reads. Therefore, it was expected that similar estimates of global repeated content in *A. albopictus* would be obtained by RE and dnaPipeTE (table 2). However, dnaPipeTE, in addition to being much faster, was also able to

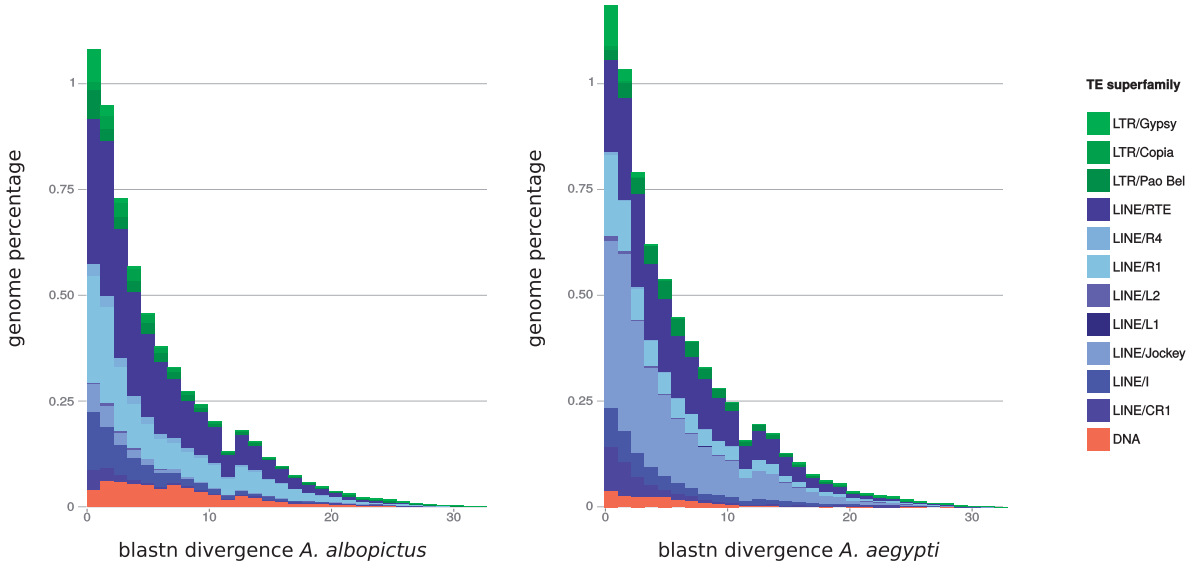



**Fig. 5.**—TE age distribution comparisons between *Aedes albopictus* (left) and *Aedes aegypti* (right). For each species, the nucleotide divergence from BLASTN is reported between a repeat read and the contig, where it matches the dnaPipeTE assembly.

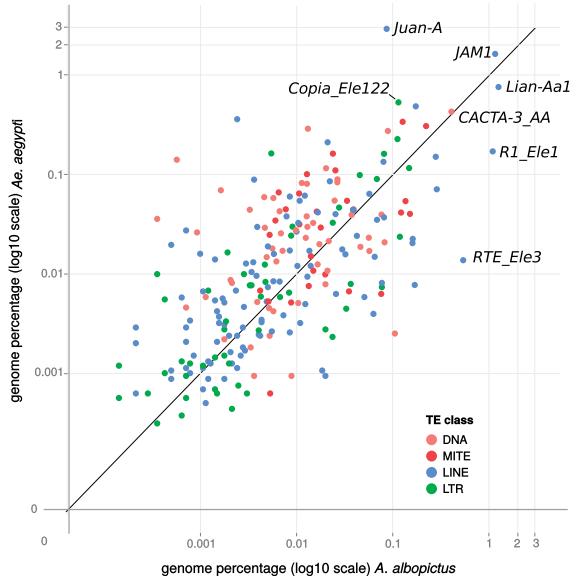



**Fig. 6.**—Comparison of the relative genome proportions of shared TE families between *Aedes albopictus* and *Aedes aegypti* in terms of genome percentage ($\log_{10}$ scale). Each dot represents a shared TE family, defined by a more similar BLAT hit between the TE family reference contig of each species. Names on the graphs correspond to the main TE annotation (from *A. aegypti*) discussed in the text.

annotate a larger fraction of TEs and to compute larger contigs. However, RE seems to more sensitively estimate the proportion of low complexity and tandem repeat sequences (data not shown).

## Discussion

### The *A. albopictus* repeatome

We report the first description of the *A. albopictus* repeatome using dnaPipeTE, a new bioinformatic pipeline for the de novo estimation, annotation, and assembly of repeatomes from raw genomic reads. We found that the total amount of repeated DNA reached 49.13% of the genome that includes at least 33.58% TEs. Taking into account that this method will underestimate low copy number TEs as well as older copies that were unable to be assembled due to mutation accumulation, our estimation should be viewed as a lower bound for the TE content of *A. albopictus*. As 7.23% of the genome is still unannotated repeats, it is possible that the TE content of *A. albopictus* ranks the largest among mosquitoes (fig. 7; Holt et al. 2002; Nene et al. 2007; Arensburger et al. 2011; Marinotti et al. 2013; Zhou et al. 2014). The large repeatome of *A. albopictus* contributes to half of its genome size, which is consistent with the observed relation between genome size and TE content (Biémont and Vieira 2004; Chénais et al. 2012). This relation exists between published genome sizes and TE content of other mosquitoes (fig. 7, $r^2 = 0.82$, $P < 0.01$).

TE families can be extremely different from each other and are classified into several subfamilies. In a given genome, some TE families are present in few copies, while others can reach hundreds of thousands of copies. In *A. albopictus*, the largest TE families in terms of genome proportion and copy numbers are LINE (non-LTR) retroelements, which harbor thousands of copies per family and represent 12.09% of the genome.

**Table 2**

Performance Comparison between dnaPipeTE and RepeatExplorer Using *Aedes Albopictus* and *Drosophila melanogaster* Samples

| | | Computing Time | Contig Number | Assembly N50 (bp) | Repeat Content Estimation | Repeat Annotation |
|---|---|---|---|---|---|---|
| *A. albopictus* | dnaPipeTE | 3 h 07 min (8 CPUs/40 Go RAM) | 8102 | 677 | 49.13% | 85.3% |
| | RepeatExplorer | 2 days 5 h 12 min (8 CPUs/16 Go RAM) | 14615 | 198 | 51.0% | 25.5% |
| *D. melanogaster* | dnaPipeTE | 0 h 40 min (8 CPUs/15 Go RAM) | 2054 | 2,590 | 18% | 98.8% |
| | RepeatExplorer | 6 h 05 min (8 CPUs/16 Go RAM) | 1352 | 287 | 16.5% | 86.1% |

Note.—Repeat annotation percentage was computed by counting the number of genomic reads receiving an annotation for each method.

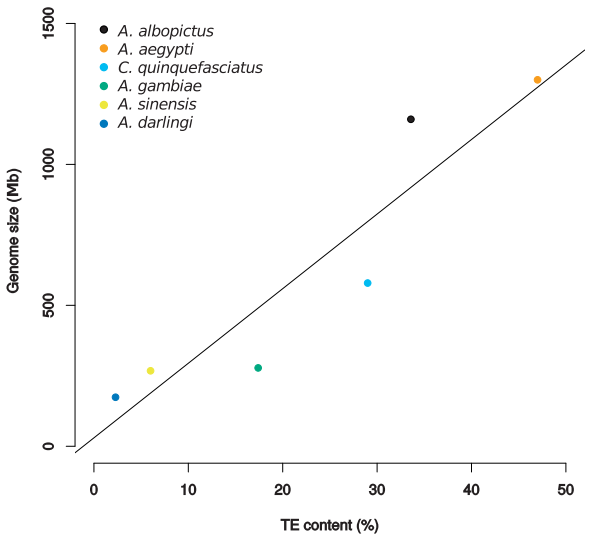



**Fig. 7.**—Linear regression of genome size over TE content in mosquitoes. Except for *Aedes albopictus*, data come from complete sequenced genomes cited in the text. ($r^2 = 0.827$, $P < 0.01$).

These LINEs represent several well-known superfamilies that have been described in mosquitoes, such as I (*Lian*, *R1*, *Loa*, and *Loner* families) and RTE (Tu et al. 1998; Biedler and Tu 2003; Boulesteix and Biémont 2005). LINEs are also found in high copy number in *A. aegypti*, where they represent 14% of the genome (Nene et al. 2007). At the class level, the most abundant class of TE is the Class II, with a majority of DNA transposons and MITEs. This feature is shared by the *A. aegypti* genome, in which Class II elements are also the most abundant repeats, comprising 20% of genome proportion, including 16% of MITEs.

## TE Dynamics and Comparison with *Aedes aegypti*

Comparison of the two related *Aedes* species highlighted a convergence in TE landscapes at the superfamily level. Both species display a similar distribution of sequenced TE reads against their contig sequences for the three TEs studied (LTR, LINEs [Class I], and Class II). In these species, Class I elements (RNA-mediated transposition) showed a right-skewed distribution, meaning that copies of each TE family share a high identity. This is typical of recent or active TE families, in which the copy number increases faster than the accumulation of mutations within the copies (Lerat et al. 2011; Staton et al. 2012). This pattern can be seen in species such as *D. melanogaster* or *An. gambiae*, in which Class I elements showed recent amplifications (Biedler and Tu 2003; Kapitonov and Jurka 2003; see also the genome analysis available online at http://repeatmasker.org/genomicDatasets/RMGenomicDatasets.html, last accessed April 13, 2015).

In both mosquito species, DNA-based transposons (Class II) are poorly represented compared with their relative genome proportion. However, this result might be explained by the removal of *A. aegypti* TE references from the library to avoid any bias toward this species in the annotation, which might have removed elements specific to the *Aedes* genus. Another explanation is that DNA transposons could belong to families with very few copies and/or result from an old invasion of the genome. Thus, our methodology, which is weaker beyond 15% divergence and for elements with few copies, could have missed old Class II elements. Ultimately, this could mean either that members of Class II are the first TEs to have invaded *Aedes* genomes or that Class I TEs are undergoing a new expansion wave.

Despite these similarities in the TE age distributions, the LINE/Jockey superfamily is different between these two species. Indeed, these elements are rare (0.04% of the blasted reads) in *A. albopictus*, where only recent copies are found. However, in *A. aegypti*, they represent half of the LINEs, and the LINE/*Juan-A* is the most abundant TE, representing 3% of the genome (Nene et al. 2007). Conversely, *A. albopictus* harbors more LINE/I elements than *A. aegypti*, and their distribution indicates a higher number of divergent copies, which suggests that their amplification in the *A. albopictus* genome could have begun earlier than in *A. aegypti* following the divergence of these two species.

The distinction between *A. albopictus* and *A. aegypti* is even more striking when observing the abundance of the TE families they share. Indeed, the abundance of TEs copies is very

different from one genome to another. This indicates that while both species share similar trends in TE class dynamics, a TE expansion occurred independently in each species. This observation could be interpreted in the ecological framework of TE dynamics and evolution (Venner et al. 2009; Linquist et al. 2013). Indeed, "ecological" factors affecting the genome, such as GC content or genome size, have been shown to be linked to TE abundance and distribution in related species (Jurka et al. 2011). Thus, inheritance of a common genome and ecosystem from an ancestor could have constrained superfamily dynamics in both species, considering either the possible interaction between TEs (identical to interspecific competition) or between TEs and the genome architecture (Venner et al. 2009; Linquist et al. 2013). However, at the family level, the spread of one TE family instead of another is not subject to ecological constraint (Jurka et al. 2011). For instance, the general pattern of a recent invasion of LTRs and LINEs in the *Aedes* species studied here can still be observed, while the specific TE families amplified in each species differ. In addition, both *A. albopictus* and *A. aegypti* are examples of species with numerous subdivided populations in their native areas (Hawley 1988; Mousson et al. 2005; Brown et al. 2014) and a relatively limited natural dispersion capability (Reiter 1996; Bellini et al. 2010; Medley et al. 2015), which increases the probability of differential TE fixation in isolated subpopulations (Jurka et al. 2011). Therefore, the sequenced individuals are only representative of the subpopulations to which they belong, and it would be interesting to compare TE family diversity at the subpopulation level with regard to intraspecific genome size variation imparted by TEs in *A. albopictus* (McLain et al. 1987; Black and Rai 1988).

### dnaPipeTE: A Novel Tool for TE Comparative Studies

Preliminary work on the *A. albopictus* repeatome led us to develop our own pipeline in order to address specific unmet needs. As the *A. albopictus* genome is especially large, we were interested in solutions using low coverage sequencing to find and quantify TEs and interspersed repeats. The most advanced software for this task previously available was RE (Novák et al. 2010), which allows the simultaneous location, quantification, and annotation of repeats from unassembled sequencing reads. However, we felt that some points could be improved by using NGS-specific tools. By using Trinity as a TE assembler on small genomic data sets, dnaPipeTE can recover larger TE contigs and can improve this step by performing multiple iterations with additional independent samples. dnaPipeTE can annotate and quantify TE families with its contigs and the number of mapped reads, while RE annotation is given only for sampled reads. Our method allowed the identification of more repeats in *A. albopictus* than RE, with a substantial decrease in computational time. As with other library-based tools, this automatic annotation should be considered with caution when working on species with very few reference libraries, where the similarities between hits might be weak and could lead to annotation errors. However, tests on model species showed that dnaPipeTE performed well in the estimation of the TE content and the proportions of the main TE families. Although it was not designed for de novo identification of new TE families, dnaPipeTE can produce full-length contigs of TEs that could be manually annotated at a later point. dnaPipeTE also provides a large amount of usable output (summary tables, graphs, sorted data sets). Finally, dnaPipeTE is the first method capable of generating a representation of TE age distribution without prior genome assembly. This analysis of course has some limitations. First, the BLAST method allows the detection of variation only from 0% to 15% divergence. Second, considering two divergent copies in a TE family, the accumulation of mutations will not be evenly distributed along the sequence; reads from a conserved protein domain will be more similar to the contig than nonfunctional regions due to selective constraints, biasing the TE age distribution toward recent divergence. In the future, the effects of these drawbacks will be reduced by the use of longer reads, which dnaPipeTE is already equipped to handle. In conclusion, this new bioinformatic pipeline, available for download at https://lbbe.univ-lyon1.fr/-dnaPipeTE-.html, allowed us to perform a fast and comprehensive analysis of TEs and repeat elements in a newly sequenced genome using NGS raw data with only 0.3× genome coverage. It allows the design of "low sequencing experiments" that reduce sequencing cost and facilitate an increase in the number of samples compared. The consistency and the robustness of dnaPipeTE also allow for comparative studies such as the one presented in this article.

Our study showed that the repeatome of *A. albopictus* is huge, encompassing 50% of the genome, and that it shares notable similarities with *A. aegypti* at the main TE order level. The intrafamily dynamics of TEs show high variation between species. Since the divergence of *A. albopictus* and *A. aegypti* 10 million years ago (Pashley and Rai 1983), TE families seemed to have evolved independently from ancestral TE ecology. These pictures of the two *Aedes* species' repeatomes could explain the large genome size variation due to repetitive DNA reported at the intraspecific level (McLain et al. 1987; Black and Rai 1988).

## Supplementary Material

Supplementary figures S1–S4, tables S1 and S2, and Material are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

## Acknowledgments

We thank the members of the Transposable Elements, Evolution, and Population team for testing dnaPipeTE and for providing suggestions and constructive discussion about

this article. We also thank Petr Novák for his availability to provide help and information about the RepeatExplorer pipeline and Murray Patterson for English revisions. This work was performed using the computing facilities of the CC LBBE/PRABI. This work was supported by the Agence Nationale de la Recherche (ANR Genemobile), the Centre National de La Recherche Scientifique, the Institut Universitaire de France, and the Federation de Recherche 41 "Bio-Environnement et Santé." C.G. received a grant from the French Ministry of Superior Education. This study was also supported by The *A. albopictus* genome sequencing project which is partly funded by the Agence Nationale de la Recherche (ANR Immunsymbart).

## Literature Cited

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol. 215:403–410.

Arensburger P, Hice RH, Wright JA, Craig NL, Atkinson PW. 2011. The mosquito *Aedes aegypti* has a large genome size and high transposable element load but contains a low proportion of transposon-specific piRNAs. BMC Genomics 12:606.

Beck CR, Garcia-Perez JL, Badge RM, Moran JV. 2011. LINE-1 elements in structural variation and disease. Annu Rev Genomics Hum Genet. 12: 187–215.

Bellini R, et al. 2010. Dispersal and survival of *Aedes albopictus* (Diptera: Culicidae) males in Italian urban areas and significance for sterile insect technique application. J Med Entomol. 47:1082–1091.

Biedler J, Tu Z. 2003. Non-LTR retrotransposons in the African malaria mosquito, *Anopheles gambiae*: unprecedented diversity and evidence of recent activity. Mol Biol Evol. 20:1811–1825.

Biémont C, Vieira C. 2004. [The influence of transposable elements on genome size]. J Soc Biol. 198:413–417.

Black WC, Ferrari JA, Sprengert D. 1988. Breeding structure of a colonising species: *Aedes albopictus* (Skuse) in the United States. Heredity (Edinb) 60(Pt 2):173–181.

Black WC, Rai KS. 1988. Genome evolution in mosquitoes: intraspecific and interspecific variation in repetitive DNA amounts and organization. Genet Res. 51:185–196.

Bonizzoni M, Gasperi G, Chen X, James AA. 2013. The invasive mosquito species *Aedes albopictus*: current knowledge and future perspectives. Trends Parasitol. 29:460–468.

Boulesteix M, Biémont C. 2005. Transposable elements in mosquitoes. Cytogenet Genome Res. 110:500–509.

Brown JE, et al. 2014. Human impacts have shaped historical and recent evolution in *Aedes aegypti*, the dengue and yellow fever mosquito. Evolution 68:514–525.

Casacuberta E, González J. 2013. The impact of transposable elements in environmental adaptation. Mol Ecol. 22:1503–1517.

Chénais B, Caruso A, Hiard S, Casse N. 2012. The impact of transposable elements on eukaryotic genomes: from genome size increase to genetic adaptation to stressful environments. Gene 509: 7–15.

Cock PJ, Fields CJ, Goto N, Heuer ML, Rice PM. 2010. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res. 38:1767–1771.

Goodier JL, Kazazian HH. 2008. Retrotransposons revisited: the restraint and rehabilitation of parasites. Cell 135:23–35.

Grabherr MG, et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29: 644–652.

Hawley WA. 1988. The biology of *Aedes albopictus*. J Am Mosq Control Assoc Suppl. 1:1–39.

Holt RA, et al. 2002. The genome sequence of the malaria mosquito *Anopheles gambiae*. Science 298:129–149.

Huang X. 1999. CAP3: a DNA sequence assembly program. Genome Res. 9:868–877.

Jurka J, Bao W, Kojima KK. 2011. Families of transposable elements, population structure and the origin of species. Biol Direct. 6:44.

Jurka J, et al. 2005. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res. 110:462–467.

Kapitonov VV, Jurka J. 2003. Molecular paleontology of transposable elements in the *Drosophila melanogaster* genome. Proc Natl Acad Sci U S A. 100:6569–6574.

Koch P, Platzer M, Downie BR. 2014. RepARK—de novo creation of repeat libraries from whole-genome NGS reads. Nucleic Acids Res. 42:e80.

Kumar A, Rai KS. 1990. Intraspecific variation in nuclear DNA content among world populations of a mosquito, *Aedes albopictus* (Skuse). Theor Appl Genet. 79:748–752.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods. 9:357–359.

Lerat E, Burlet N, Biémont C, Vieira C. 2011. Comparative analysis of transposable elements in the melanogaster subgroup sequenced genomes. Gene 473:100–109.

Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22: 1658–1659.

Linquist S, et al. 2013. Distinguishing ecological from evolutionary approaches to transposable elements. Biol Rev Camb Philos Soc. 88: 573–584.

Marinotti O, et al. 2013. The genome of *Anopheles darlingi*, the main neotropical malaria vector. Nucleic Acids Res. 41:7387–7400.

McLain DK, Rai KS, Fraser MJ. 1987. Intraspecific and interspecific variation in the sequence and abundance of highly repeated DNA among mosquitoes of the *Aedes albopictus* subgroup. Heredity (Edinb) 58: 373–381.

Medley KA, Jenkins DG, Hoffman EA. 2015. Human-aided and natural dispersal drive gene flow across the range of an invasive mosquito. Mol Ecol. 24:284–295.

Modolo L, Lerat E. 2014. Identification and analysis of transposable elements in genomic sequences. In: Poptsova MS, editor. Genome analysis: current procedures and application. Norfolk (UK): Caister Academic Press. p. 165–181.

Mousson L, et al. 2005. Phylogeography of *Aedes (Stegomyia) aegypti* (L.) and *Aedes (Stegomyia) albopictus* (Skuse) (Diptera: Culicidae) based on mitochondrial DNA variations. Genet Res. 86:1–11.

Nene V, et al. 2007. Genome sequence of *Aedes aegypti*, a major arbovirus vector. Science 316:1718–1723.

Novák P, Neumann P, Macas J. 2010. Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. BMC Bioinformatics 11:378.

Pashley DP, Rai KS. 1983. Comparison of allozyme and morphological relationships in some *Aedes* (*Stegomyia*) mosquitoes (Diptera: Culicidae). Ann Entomol Soc Am. 76:388–394.

Rao PN, Rai KS. 1987. Inter and intraspecific variation in nuclear DNA content in *Aedes* mosquitoes. Heredity (Edinb) 59:253–258.

Rebollo R, Horard B, Hubert B, Vieira C. 2010. Jumping genes and epigenetics: towards new species. Gene 454:1–7.

Reiter P. 1996. [Oviposition and dispersion of *Aedes aegypti* in an urban environment]. Bull Soc Pathol Exot. 89:120–122.

Staton SE, et al. 2012. The sunflower (*Helianthus annuus* L.) genome reflects a recent history of biased accumulation of transposable elements. Plant J. 72:142–153.

Tange O. 2011. GNU parallel: the command-line power tool. ;login USENIX Mag. 3:42–47.

Tu Z, Isoe J, Guzova JA. 1998. Structural, genomic, and phylogenetic analysis of *Lian*, a novel family of non-LTR retrotransposons in the yellow fever mosquito, *Aedes aegypti*. Mol Biol Evol. 15:837–853.
Vela D, Fontdevila A, Vieira C, García Guerreiro MP. 2014. A genome-wide survey of genetic instability by transposition in *Drosophila* hybrids. PLoS One 9:e88992.
Venner S, Feschotte C, Biémont C. 2009. Dynamics of transposable elements: towards a community ecology of the genome. Trends Genet. 25:317–323.
Wicker T, et al. 2007. A unified classification system for eukaryotic transposable elements. Nat Rev Genet. 8:973–982.
Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18:821–829.
Zhou D, et al. 2014. Genome sequence of *Anopheles sinensis* provides insight into genetics basis of mosquito competence for malaria parasites. BMC Genomics 15:42.
Zytnicki M, Akhunov E, Quesneville H. 2014. Tedna: a transposable element de novo assembler. Bioinformatics 30:2656–2658.

# Supplementary material

*De-novo* assembly and annotation of the Asian tiger mosquito (*Aedes albopictus*) repeatome from raw genomic reads with dnaPipeTE and comparative analysis with the yellow fever mosquito (*Aedes aegypti*)

Clément Goubert, Laurent Modolo, Cristina Vieira, Claire Valiente-Moro Patrick Mavingui, and Matthieu Boulesteix

## Contents

# 1 Test of dnaPipeTE efficiency

To assess the method performances and boundaries, we tested dnaPipeTE on a set of reference genomes for which both a fully assembled genome and suitable NGS sequences were available. We then compared the genome proportion of the main TE families and the TE age distribution generated either with dnaPipeTE or with RepeatMasker for assembled genomes. RepeatMasker analyses were already performed by the A. Smit team (Institute for Systems Biology) and are fully available online at `http://repeatmasker.org/genomicDatasets/RMGenomicDatasets.html`

## 1.1 Datasets

When available, we used the same strain in dnaPipeTE analysis and for the assembled one. We downloaded NGS datasets from the Short Read Archive (http://www.ncbi.nlm.nih.gov/sra) and cleaned the datasets using fastx-toolkit with the parameters described in the Material and Methods section. For the smaller genomes, we used an initial sample size of 0.25X coverage, while we used 0.1X coverage for the larger ones (*Ae. aegypti* and *Homo sapiens*). We used different sample sizes because preliminary results showed that increasing sample size did not substantially increase the total number of repeat sequences found, while it exponentially increased the computation time. We only used the R1 end of each file (single end) for dnaPipeTE runs. The following table (Suppl. Table S1) summarises the dataset specifications.

## 1.2 Conclusions

In most cases (see Suppl. Figures S1 and S3), dnaPipeTE was able to find most of the TEs and other repeat classes described, using NGS datasets. Globally, the estimation of the total number of repeats using dnaPipeTE appears relevant with regards to the estimation made from the assembled genomes. Estimation of the TE content was very good for samples using exactly the same strain in both analyses (i.e., *D. melanogaster* and *C. elegans*). Estimations were also very good for *C. intestinalis* and *A. gambiae* for which the strains were different or unidentified; we noticed that in both species, dnaPipeTE identified new repeats.

Looking at the results from *G. aculeatus*, we found many more repeats with dnaPipeTE than with RepeatMasker analysis of the assembled genome. New annotations mostly came from other fishes; thus, we are quite confident that they are not false positives. However, it is possible that the current *G. aculeatus* did not include all of the repeats, depending on the sequencing and assembly method used. Although they came from the same place (Bear Paw Lake), the NGS-sequenced sample is divergent from the reference sequenced individual.

In *Ae. aegypti*, we note that the total numbers of Class II repeats (MITEs and DNA) were close (19% vs 24%). However, dnaPipeTE found fewer MITEs than expected. MITEs are short TEs without coding sequences and are derived from DNA transposons. It is thus possible that most of them have been identified as DNA in the absence of better available annotation.

From the TE landscape estimations, we were able to show that our method allows us to catch variation in TE age distribution up to at least 15% divergence. Beyond this threshold, the blastn method fails to match reads with more divergent contigs, and thus those TEs will be dropped from the report. However, we can clearly distinguish characteristic shapes between models that fit

with the fully assembled genomic TE landscapes that are available at http://repeatmasker.org/genomicDatasets/RMGenomicDatasets.html. In addition, we can note that the first bin in those graphs (0 to 1% divergence) is inflated compared to RepeatMasker analysis of fully assembled genomes. This issue is discussed in the paper.

### 1.3 Case of an "old" repeatome

To test the performance of dnaPipeTE with genomes that have old TE families with high divergence between copies, we tested the pipeline on the human genome, which is representative of such old repeatomes (Suppl. Figure S4).

For the human genome, dnaPipeTE did not manage to accurately estimate the abundance of repeated content. While the relative proportion of each TE classes is well estimated, we found that it only represented half of its actual size. This is certainly due to the particular profile of the TE age in the human genome, where TEs are mostly ancient. Thus, if there is less identity between reads from different copies of the same TE family, the assembly will fail to find the older families. This is the main limitation of this method with regard to low coverage datasets.

## 2 Sample size choice for *Ae. albopictus* and the interest of multiple iteration

To maximise the N50 of assembly while minimising inclusion of non-repetitive genome content, we tested dnaPipeTE with different sample sizes and Trinity iterations (see Materials and Methods). We found that in *Ae. albopictus*, the best compromise was to choose a combination of two iterations and a sample size of 0.1X. In Suppl. Figure S2, each combination of Trinity iteration and sample size was tested two times, except for 0.1X and 0.17X, which have three repetitions each.

In *D. melanogaster*, after the first Trinity assembly on a 0,25X sample, the N50 was 1342 bp for 1302 contigs. Then, after adding a new sample of 0.25X to the assembled reads for the second iteration, the N50 rose to 2054 bp with 2590 contigs.

## 3 Supplementary data

Supplementary data 1 : annotated full-length contigs (dnaPipeTE_full_lengths_TE_albo.fasta)
Supplementary data 2: annotated partial contigs (dnaPipeTE_partial_TE_albo.fasta)

# 4 Supplementary Tables

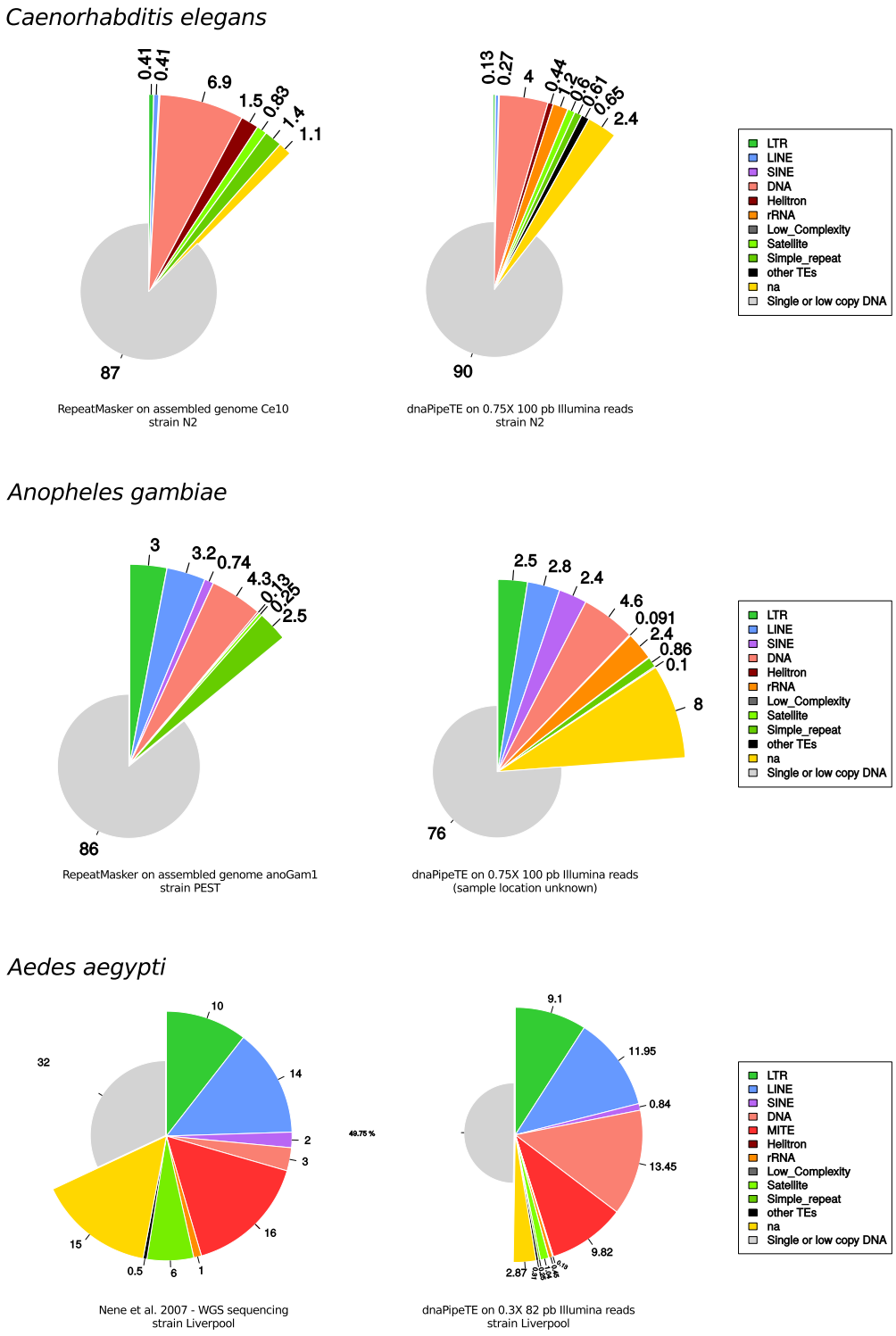
Suppl. Table S 1: Species and dataset used for dnaPipeTE tests. Genomes sizes are taken from whole genome assemblies and can be found at http://repeatmasker.org/. Unless otherwise stated, datasets used for dnaPipeTE are the R1 reads from 101 bp Illumina HiSeq2000 sequencing

| **Species** | **Genome size** | **Assembled strain** | **RepeatMasker Analysis** | **dnaPipeTE strain** | **dnaPipeTE samples size (assembly + blast sample)** | **NCBI SRA archive** |
|---|---|---|---|---|---|---|
| *Drosophila melanogaster* | 162 Mbp | *w1118* | *http://repeatmasker.org /species/dm.html* | *w1118* | 2x 0,25X + 0,25 X | SRR988075[a] |
| *Anopheles gambiae* | 263 Mbp | PEST | *http://repeatmasker.org /species/anoGam.html* | *Unknown* | 2x 0,25X + 0,25 X | ERR554052[b] |
| *Caenorhabditis elegans* | 100 Mbp | N2 | http://repeatmasker.org/ species/ce.html | N2 | 2x 0,25X + 0,25 X | DRR008444[c] |
| *Ciona intestinalis* | 141 Mbp | HK | http://repeatmasker.org/ species/ci.html | T | 2x 0,25X + 0,25 X | DRR018354[d] |
| *Gasterosteus aculeatus* | 447 Mbp | Bear Paw | http://repeatmasker.org/ species/gasAcu.html | Bear Paw | 2x 0,25X + 0,25 X | SRR070080[e] |
| *Homo sapiens* | 3 Gbp | hg38 | http://repeatmasker.org/ species/hg.html | NA18912 | 2x0,1X + 0,1X | SRR350153[f] |
| *Aedes aegypti* | 1,3 Gbp | Liverpool | -- | Liverpool | 2x0,1X + 0,1X (82 bp Illumina Hiseq2000 reads) | SRR871496[g] |

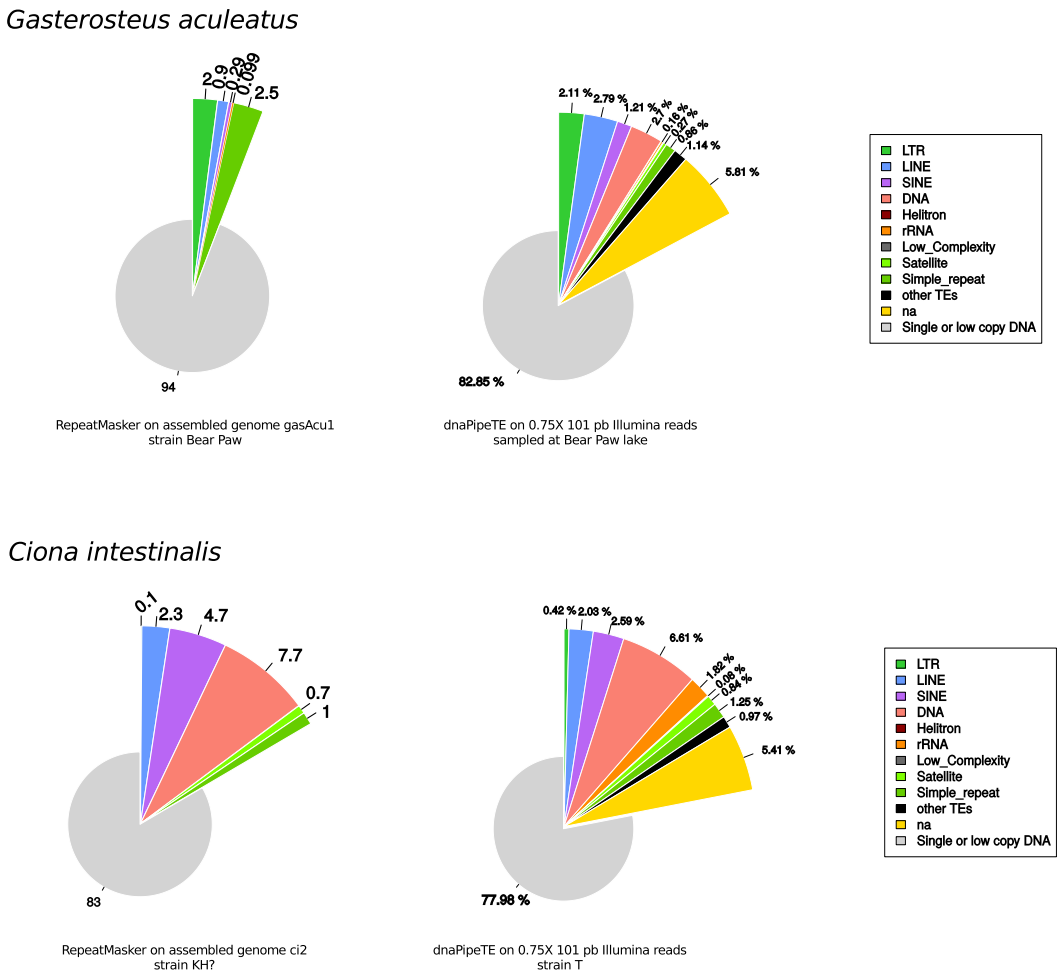
[a] Beijing Institute of Genomics, CAS, 2013
[b]Anopheles Genome Variation Project, 2014
[c]Center for Genetic Resource Information, Comparative Genomics Laboratory, National Institute of Genetics, Research Organization of Information and Systems, 2014
[d]Marine Genomics Unit, Okinawa Institute of Science and Technology
[e]Broad Institute, 2006
[f]1000 Genomes Project, 2008
[g]Virginia Tech, 2013

Suppl. Table S 2: Blat results of the *Ae. albopictus* assembled transcriptome on dnaPipeTE contigs (blast format). *The table file is:* "Transcriptome_to_dnaPipeTE_contigs_best80pb_80pcId_bestBitScore.blastformatout" Column: 1 Transcriptome contig (query); 2 dnaPipeTE contigs (target); 3 Percentage indentity 4 Alignment size (bp); 5 # mismatches; 6 # gap opening; 7 Query start; 8 Query end; 9 Subject start; 10 Subject end; 11 E-value; 12 Bit Score

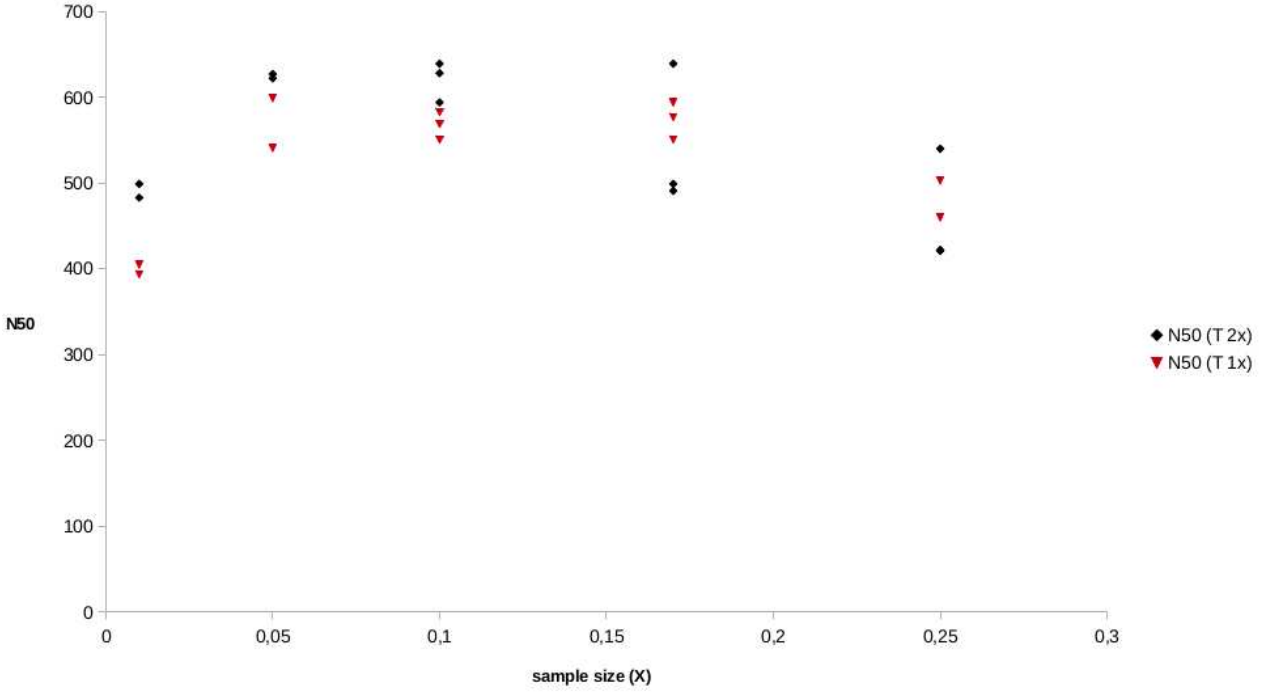
# 5 Suppelementary Figures




Suppl. Figure S 1: Estimation of the repeated content with dnaPipeTE and a comparison with whole assembled genome analysis. Pairs of pie charts summarise the overall number of repeat classes either using RepeatMasker on the whole assembled genome (left) [except for *Ae. aegypti*, data from the TE content analysis performed by Nene et al. 2007] or dnaPipeTE on single-end Illumina reads (right). Values are given as percentage of the genome content. *na*: no annotation found



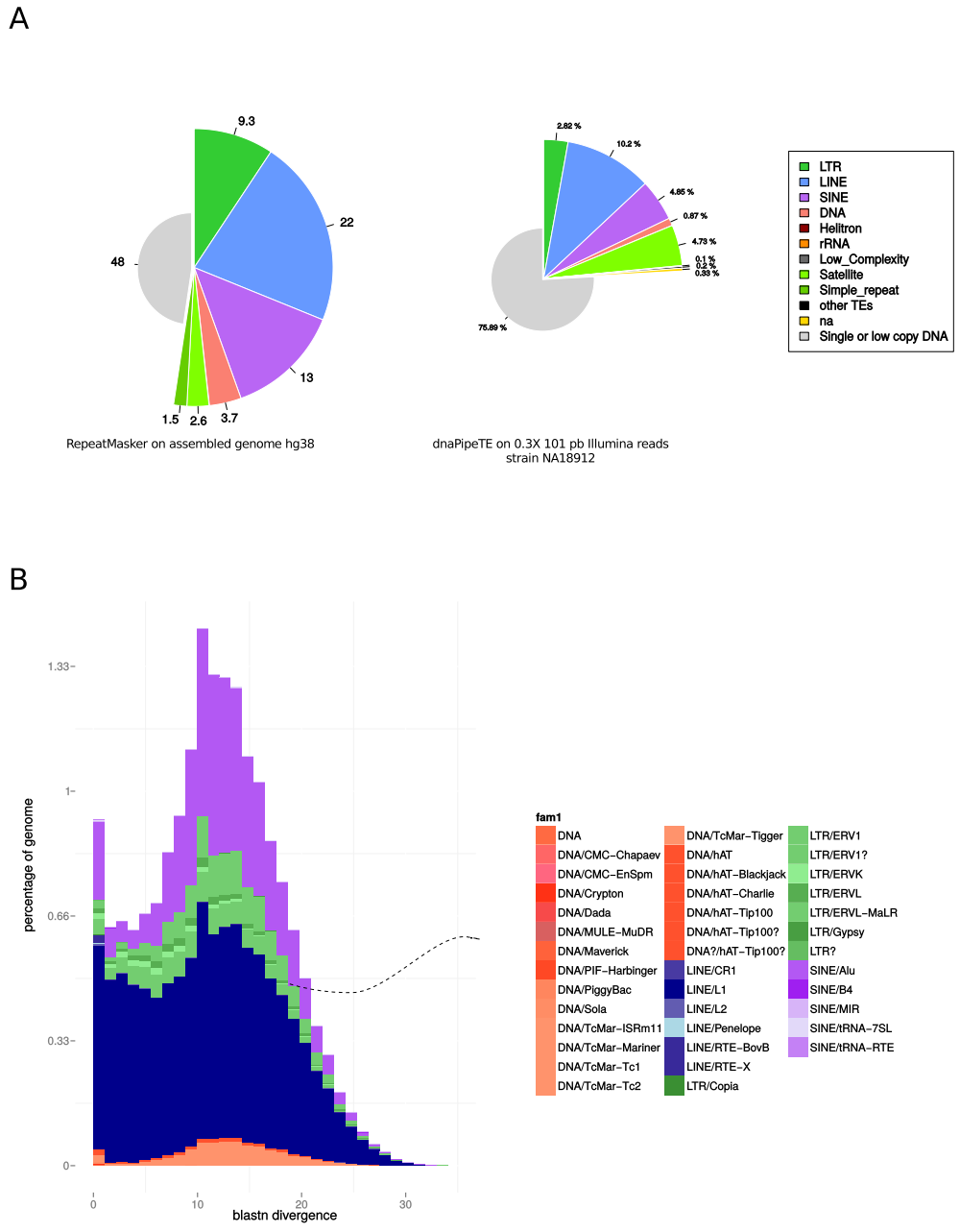

Suppl. Figure S 1: Continued. Estimation of the repeated content with dnaPipeTE and a comparison with whole assembled genome analysis. Pairs of pie charts summarise the overall number of repeat classes either using RepeatMasker on the whole assembled genome (left) [except for *Ae. aegypti*, data from the TE content analysis performed by Nene et al. 2007], either dnaPipeTE on single end Illumina reads (right). Values are given in percentage of the genome content.*na*: no annotation found




Suppl. Figure S 2: Assembly N50 after the first (T 1x) and the second (T 2x) Trinity iteration in dnaPipeTE for *Ae. albopictus*, according to sample size. For T 2x, two samples of the same size were used successively, according to the description provided in Materials and Methods

Suppl. Figure S 3: TE age distribution (TE landscapes) from blastn divergence between reads and their dnaPipeTE contigs. Bar plots represent the total number of reads from TEs (as a genome percentage) according to their divergence from the matching dnaPipeTE contig, using blastn. Dashed black lines represent the shape of the distribution observed using RepeatMasker on assembled genomes.

*Homo sapiens*




Suppl. Figure S 4: A. The estimation of repeated content with dnaPipeTE and a comparison with whole assembled genome analysis of the *Homo sapiens* genome. B. TE age distribution (TE landscapes) from blastn divergence between reads and their dnaPipeTE contigs. The dashed black line represents the shape of the distribution observed using RepeatMasker on the hg38 genome assembly

# Chapitre 3

# Recherche sans *a priori* de traces d'adaptation sur le génome du moustique tigre

*"The most exciting phrase to hear in science, the one that heralds new discoveries, is not 'Eureka!' but 'That's funny...'"*

– Isaac Asimov

# Avant propos

À la faveur des travaux précédents, nous disposions des ressources nécessaires à la réalisation du scan génomique entre populations tropicales présentes dans l'aire d'origine et populations tempérées, invasives en Europe.

L'analyse du répétome d'*Ae. albopictus* a fourni des séquences de référence pour 5 familles d'ET particulièrement abondantes, leur nombre de copies étant estimé entre 500 et 4000 par génome selon les familles d'ET par la méthode dnaPipeTE. Nous avons dans un premier temps procédé à l'amplification par PCR de ces ET chez quelques individus, afin de valider leur présence et la spécificité des amorces. Pour cela, un protocole de TD identique (pour sa partie pré-séquençage) à celui présenté dans l'article qui suit à été utilisé, et les produits de PCR ont été clonés et séquencés.

Dans un second temps, un financement de la fédération de recherche 41 « Bio-environnement-santé » nous a permit de réaliser une expérience pilote, de séquençage à haut-débit des produits de TD. Pour cela, nous avons construit une banque de séquences correspondante aux insertions de trois familles d'ET (les familles L2B, IL1 et RTE5 utilisées dans l'article), pour 4 individus provenant de deux populations du Vietnam. Ces échantillons ont été séquencés sur plateforme 454 Junior (Roche), et nous avons obtenu 100 000 lectures (non pairées, de tailles comprises entre 200 et 500pb), chacune d'elle devant correspondre à l'insertion d'une copie d'ET. Ces données nous ont par ailleurs permis de tester les outils bio-informatiques utilisés ensuite lors de l'expérience à grande échelle. Grâce à ce pilote, nous avons pu reconstituer 1091 locus dont seulement 2 étaient détectés comme présentes chez les 4 individus, pour un nombre de locus moyen de 324 par individu (les résultats détaillés de cette analyse sont présentés en Annexe 2). Ces résultats permettaient d'être optimistes quant à ceux obtenus à plus grande échelle, notamment car nous comptions séquencer 140 individus avec une couverture environ 30 fois supérieure (100 000 000 de paires de séquences), ce qui permettait de prévoir un nombre plus important de locus détectés mais aussi partagés par un nombre plus important d'individus.

L'expérience finale comprenait donc 140 moustiques, appartenant à 3 populations vietnamiennes, 3 françaises et une espagnole, échantillonnée près de Barcelone. L'article présenté dans ce chapitre rapporte en détail les différentes étapes du protocole de génotypage à haut débit, et l'analyse de scan génomique qu'il a été possible de réaliser par la suite.

Parmi les résultats principaux, nous avons obtenus plus de 120 000 locus polymorphes sur lesquels réaliser nos analyses. L'étude de la structure des populations à confirmé le patron connu concernant l'absence de différenciation par la géographie, et l'importante diversité génétique retrouvée entre les individus au sein des populations. Ces résultats nous ont conduits à utiliser le modèle démographique en îles de la méthode Bayescan, qui envisage que les différentes populations puissent échanger des migrants à des taux diffé-

rents, cependant sans structure hiérarchique. L'analyse des résultats du scan génomique nous a permis d'identifier 92 locus outliers, dont la plupart sont dus à des fortes fréquences d'insertion en Europe, un patron original, possible signe d'une adaptation récente au sein de ces environnements.

# High Throughput Transposable Elements insertion polymorphism genotyping reveals adaptive evolution toward temperate environment in the invasive Asian tiger mosquito *Aedes albopictus*

Clément Goubert[1], Hélène Henri[1], Guillaume Minard[2,3], Claire Valiente Moro[2], Patrick Mavingui[2,4], Cristina Vieira[1], and Matthieu Boulesteix[1]

[1]Laboratoire de Biométrie et Biologie Evolutive, UMR CNRS 5558, INRIA, VetAgro Sup, Université Claude Bernard Lyon 1, Villeurbanne, France
[2]Microbial Ecology, UMR CNRS 5557, USC INRA 1364, VetAgro Sup, FR41 BioEnvironment and Health, Université Claude Bernard Lyon 1, Villeurbanne, France
[3]Metapopulation Research Group, Department of Biosciences, University of Helsinki, Helsinki, Finland
[4]Université de La Réunion, UMR PIMIT, CNRS 9192, INSERM 1187, IRD 249

**Abstract**

Invasive species represent unique opportunities to evaluate the role of local adaptation during colonisation of new environments. The Asian tiger mosquito *Aedes albopictus* is a major vector of Dengue and Chikungunya viruses and has spread throughout the world from south-eastern Asia in less than forty years. Its presence in both temperate and tropical environments has often been considered to be the reflect of a great "ecological versatility" found in this species. However, few studies have been conducted to asses the role of adaptive evolution in the ecological succes of *Ae. albopictus*. We thus performed a genomic scan between tropical native populations of Vietnam and temperate invasive populations in Europe to search for potential signatures of selection leading to local adaptation. To do so, we developed a Transposon Diplay method based on the high throughput genotyping of insertions of 5 families of Transposable Elements (TE) highly repeated in the *Ae. albopictus* genome. This led us to obtain from a hundred of field collected mosquitoes more than 120 000 polymorphic loci, without the need of a reference genome. While the largest part of our markers revealed a virtual absence of structure between bio-geographic areas, genome scan analyses revealed 92 outlier loci with a high level of differentiation between temperate and tropical populations. In addition, we found that a significant majority of these outliers were due to high insertion frequencies of those markers among European temperate populations, and we concluded that events of adaptive evolution could have recently occurred in the genome of temperate populations.

**Keywords:** *Invasive species, local adaptation, genome scan, Transposon Display, high-throughput sequencing*

## Introduction

Biological invasions, in spite of their dramatic impacts at ecological and societal levels, represent unique opportunities to study fast evolutionary changes such as adaptive evolution. Indeed, settlement into a novel area represent a biological challenge that invasive species have successfully overcome. The underlying processes could be thus studied at the molecular level, particularly to gather empirical knowledge about the genetics of invasions, a field of study that has produced extensive theoretical assumptions, but for which there is still little evidence in nature (Colautti et Lau, 2015). Some of the main concerns are to disentangle the effect of neutral effects during colonization, such as founder events or allele surfing at the migration front, from adaptive evolution (local adaptation, Colautti et Lau, 2015; Lande, 2015; Peischl et Excoffier, 2015).

Adaptation could arise through the appearance of a new beneficial mutation, the spread of favorable allele from standing genetic variation or from hy-

bridization in the introduction area (Handley *et al.*, 2011; Bock *et al.*, 2015; Colautti et Lau, 2015). Detection of the footprint of natural selection is however dependent on the availability of informative genetic markers, meaning that they should provide a substantial coverage of the genome to allow selection scans and be easily and confidently scored across many individuals. Unfortunately, invasive species are often not model species, making the development of a reliable and efficient marker challenging.

The Asian tiger mosquito, *Aedes* (*Stegomya*) *albopictus* (Diptera:Culicidae) is currently one of the most threatening invasive species (Invasive Species Specialist Group); originating from South-Eastern Asia, this species is one of the primary vectors of Dengue and Chikungunya viruses, and is also involved in the transmission of other arboviruses and parasites (Paupy *et al.*, 2009). *Ae. albopictus* has now settled in every continent except Antarctica, and is found both under tropical and temperate climates (Bonizzoni *et al.*, 2013). While this species is supposed to originate from rain forests of South-Eastern Asia (Hawley, 1988), the native area of *Ae. albopictus* encompasses contrasted environments including temperate regions of Japan and China, offering a large potential of fit toward newly colonized environments. For example, the induction of photoperiodic diapause in temperate areas, that has a genetic basis in *Ae. albopictus* (Hawley *et al.*, 1987; Hanson et Craig, 1994), is decisive to ensure invasive success in Europe or Northern America. Such a trait appears governed by a "genetic toolkit" involving numerous genes and metabolic networks (Poelchau *et al.*, 2013a,b,c) for which however the genetic polymorphism between diapausing and non-diapausing strains remains to be elucidated. In addition, the colonization of new areas that look similar at a first glance can still involve *de novo* adaptation: indeed, even environments that share climatic variables are not necessarily similar if edaphic and biotic interactions are considered (Colautti et Lau, 2015). Hence, this suggests that whatever the native and settled environment, it could be possible to find evidence of adaptive evolution in invasive populations of *Ae. albopictus*.

In order to better understand the invasive success of this species, we genotyped 140 field individuals, collected from 3 Vietnamese (native tropical area) and 5 European (invasive temperate area) populations, aiming to identify genomic region involved in local adaptation. To do so, we developed new genetic markers, based on high throughput genotyping of the insertion polymorphism of Transposable Elements (TEs). Such genetic elements represent at least one third of the genome of *Ae. albopictus* and include recently active families that could reach thousands of copies in one genome (Goubert *et al.*, 2015). In mosquitoes, TEs have been shown to be powerful markers both for population structure analysis (Biedler et Tu, 2003; Boulesteix *et al.*, 2007; Esnault *et al.*, 2008; Santolamazza *et al.*, 2008) and genome scans (Bonin *et al.*, 2008).

Amplification of TEs insertions is particularly efficient to obtain a large number of genetic markers throughout one genome, especially if few genomic resources are available, which is the case for the Asian tiger mosquito. We hypothesize that some TE insertion sites could be located at the neighborhood of targets of natural selection and thus reach high level of differentiation between native and invasive populations if selective sweeps occurred during local adaptation. In addition, some TEs could also insert near or inside coding regions and thus could be directly involved in environmental adaptation (Casacuberta et González, 2013), eventually contributing to the success of invasive species (Stapley *et al.*, 2015).

To distinguish between neutral demographic effects and adaptive evolution, we first performed population genetics analyses to reveal the global genetic structure of the studied populations. We then performed a genomic scan for selection using the Bayescan software (Foll et Gaggiotti, 2008), whose model both fitted the observed genetic structure (non-hierarchical islands model) and that is able to handle dominant genetic markers such as the insertion polymorphism of TEs. Thanks to the successful development of our TE markers, we identified a hundred of candidate loci under directional selection, for which the looming availability of an annotated genome in *Ae. albopictus* would shed light on the genomic evolution underlying the invasive success of the species.

## Material and Methods

### Biological samples

A total number of 140 flying adult females *Ae. albopictus* were collected in the field at eight sampling sites in Europe and Vietnam during the sum-

mers 2012 and 2013 (Figure 1). Individuals were either sampled using a single trap or using aspirators through the sampling site within a 50 meters radius. When traps were used, live mosquitoes were collected after a maximum of 2 days. Details about sampling sites and collection method is detailed in table 1. NCE and CGN are very close locations and correspond to the putative introduction point of *Ae. albopictus* in France (Delaunay *et al.*, 2009), BCN has been the first report of this species in Spain (Aranda *et al.*, 2006).

## High throughput Transposon Display (TD) genotyping

Insertion polymorphism of five transposable elements families: I Loner Ele_1 (IL1), Loa Ele_2B (L2B), RTE4, RTE5 and Lian 1 identified by Goubert *et al.* (2015) in *Ae. albopictus* were characterized. These TE families were chosen according to their high estimate of copy number (from 513 to 4203 cp), high identity between copies, and a "copy and paste" mode of transposition (all these TEs are non-LTR Class I retrotransposon). The protocol was made up combining methods from previous studies (Munroe et al, 1994; Roy; Esnault et al. 2008; Akkouche et al. 2012; Carnelossi et al., 2014) with high throughput Illumina sequencing of TD products (Figure **??**).

**DNA Extraction and TD adapted ligation.** Total DNA was extracted from whole adult bodies following the Phenol-Chloroform protocol described by (Minard *et al.*, 2015). Individual extracted DNA (≈ 75ng) was then used for enzymatic digestion in a total volume of 20 $\mu$L, with HindIII enzyme (10U/$\mu$L) and buffer R (Thermo Scientific) for 3 hours at 37 °C. The enzyme was inactivated at 80 °C for 20 minutes. TD adapters were set up hybridizing Hindlink with MSEB oligonucleotides (100$\mu$M, see Table 2) in 20X SSC and 1M Tris in a total volume of 333$\mu$L after 5mn of initial denaturation at 92 °C and 1h at room temperature for hybridization of the two parts. Once ready, TD adapters were then ligated to 20 $\mu$L of the digested DNA mixing 2$\mu$L of TD adapter with 10U T4 ligase and 5X buffer (Fermentas) in a final volume of 50$\mu$L for 3 hours at 23 °C.

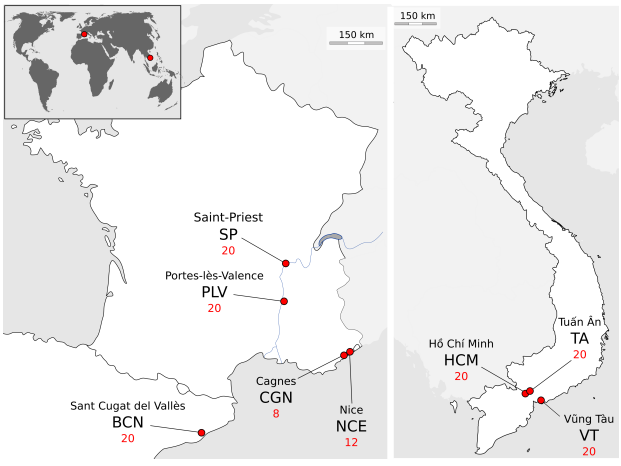



Figure 1 – Sampling sites of *Ae. albopictus* in Europe and Vietnam

**Library construction: PCR 1 and pooling per individual.** For each individual, and for each of the 5 TE families, TE insertions were amplified by PCR (PCR 1) in a Biorad Thermal Cycler (either C1000 or S1000), in a final volume of 25$\mu$L. Mixture contained 2$\mu$L of digested-ligated DNA with 1$\mu$L dNTPs (10mM), 0,5$\mu$L TD-adapter specific primer (LNP, 10$\mu$M, see primers table) and 0,5 $\mu$L of TE specific primer (10$\mu$M), 1U AccuTaq polymerase (5U/$\mu$L) with 10X buffer and Dimethyl-Sulfoxyde (Sigma). Amplification was performed as follows: denaturation at 98 °C for 30 seconds then 30 cycles including 94 °C for 15 seconds, hybridization at 60 °C for 20 seconds and elongation at 68 °C for 1 minute; final elongation was performed for 5 minutes at 68 °C. For L2B and RTE5 TEs, a nested PCR was performed in order to increase specificity in the same PCR conditions using internal forward TE primers and LNP (Table 2). PCR 1 primers include a shared tag sequence that was used for hybridization of the individual indexes by PCR 2.

For each TE, three independents PCR 1 were performed from the same digestion product. PCR 1 products (3 PCR * 5 TE per individual) were then purified using volume to volume Agencout AMPure XP beads (20 $\mu$L PCR 1 + 20 $\mu$L beads) and eluted in 30$\mu$L Resuspension buffer. After nanodrop quantification, equimolar pools containing the 3*5 PCR products per individual were made using Tecan EVO200 robot. Individual pools were then size selected for fragment ranging from 300 to 600 bp using Agencout AMPure XP beads as follow: first magnetic beads were diluted in H2O with a ratio of 1: 0.68 then add to 0.625 X PCR products in order to exclude long fragments. A second purification

Table 1 – Sampling of *Ae. albopictus*. N refers to the number of individuals from this sampling site used in our analysis. * This population referred as BG in (Minard *et al.*, 2015).

| Country | sampling site | acronym | GPS coordinates | method | N | year |
|---|---|---|---|---|---|---|
| Spain | Sant Cugat del Vallès | BCN | N41 ° 29'7" E2 ° 3'25" | aspirator | 20 | 2013 |
| France | Cagne-sur-mer | CGN | N43 ° 64'34" E7 ° 09'12" | BG sentinel | 8 | 2012 |
| France | Nice | NCE | N43 ° 41'60" E7 ° 17'33" | aspirator | 12 | 2012 |
| France | Portes-lès-Valence | PLV | N44 ° 52'8" E4 ° 52'9" | Mosquito Magnet | 20 | 2012 |
| France | Saint-Priest | SP | N45 ° 41'49" E4 ° 58'50" | Mosquito Magnet | 20 | 2012 |
| Vietnam | Hô-Chi-Minh | HCM | N10 ° 47'19" E106 ° 42'19" | aspirator | 20 | 2012 |
| Vietnam | Tuan Ân* | TA | N10 ° 57'13" E106 ° 41'50" | aspirator | 20 | 2012 |
| Vietnam | Vung Tào | VT | N10 ° 22'26" E107 ° 4'13" | aspirator | 20 | 2012 |

was performed using a non-diluted bead: DNA ratio of 1:8.3 to exclude small fragments.

**Library construction: PCR 2 (multiplexing) and purification** Single multiplexing was performed using home made 6 bp index (supplementary table 1), which were added to the R primer (see Table 2) during a second PCR with 12 cycles in ABI 2720 Thermal Cycler. Mixture contained 15ng PCR products, 1$\mu$l of dNTPs (10mM), 0.5$\mu$l MTP Taq DNA Polymerase (5U/$\mu$l, Sigma), 5$\mu$l 10X MTP Taq Buffer and 1.25$\mu$l of each tagged-primer (20$\mu$m) in a final volume of 50$\mu$l. Amplification was performed as follows: denaturation at 94 ° C for 60 seconds then 12 cycles including denaturation at 94 ° C for 60 seconds, hybridization at 65 ° C for 60 seconds and elongation at 72 ° C for 60 seconds; final elongation was performed for 10 minutes at 72 ° C. PCR 2 products were purified using Agencout AMPure XP beads: DNA ratio of 1:1.25 to obtain librairies.

**Pool and paired-end Illumina sequencing** After nanodrop quantification, librairies were equimolary pooled using Tecan EVO150 robot and the pool was then quantified by QPCR using the KAPA Library Quantification Kit to obtain an accurate quantification. Finally the pool was

Table 2 – Adapter sequences and PCR1 primers used for TD. Lowercase are universal tag used for annealing of individual indexes during PCR2. The bold part of Hindlink is specific of the *HindIII* cut site. *from Roy et al. 1999

| Primer name | Sequence 5'-3' |
|---|---|
| Hindlink | **AGCT**GAAGGAGAGGACGCTGTCTGTCGAAGG |
| MSEB* | AGCGAATTCGTCAACATAGCATTTCTGTCCTCTCCTTC |
| IL1 F | ctttccctacacgacgctcttccgatctGGCTTCCACCCATTACTAACAG |
| L2B (1/2) F | GGATCAGGTGTTACATCAACCAT |
| L2B (nested 2/2) F | ctttccctacacgacgctcttccgatctACAGCTCAATTGTGACAGGA |
| RTE5 (1/2) F | AGCGAAAACAATATTCAGCAGAG |
| RTE5 (nested 2/2) F | ctttccctacacgacgctcttccgatctATCTACGATCCCTACACGTTCG |
| RTE4 F | ctttccctacacgacgctcttccgatctGTGCGGAACTGGAGACAAAC |
| Lian1 F | ctttccctacacgacgctcttccgatctTCCTTCCTTTTCCCTCAGGT |
| LNP (TD adapter) R | ctttccctacacgacgctcttccgatctGAATTCGTCAACATAGCATTTCT |
| PCR2 F | AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGAC |
| PCR2 R | CAAGCAGAAGACGGCATACGAGAT-index-GTGACTGGAGTTCAGACGTGT |

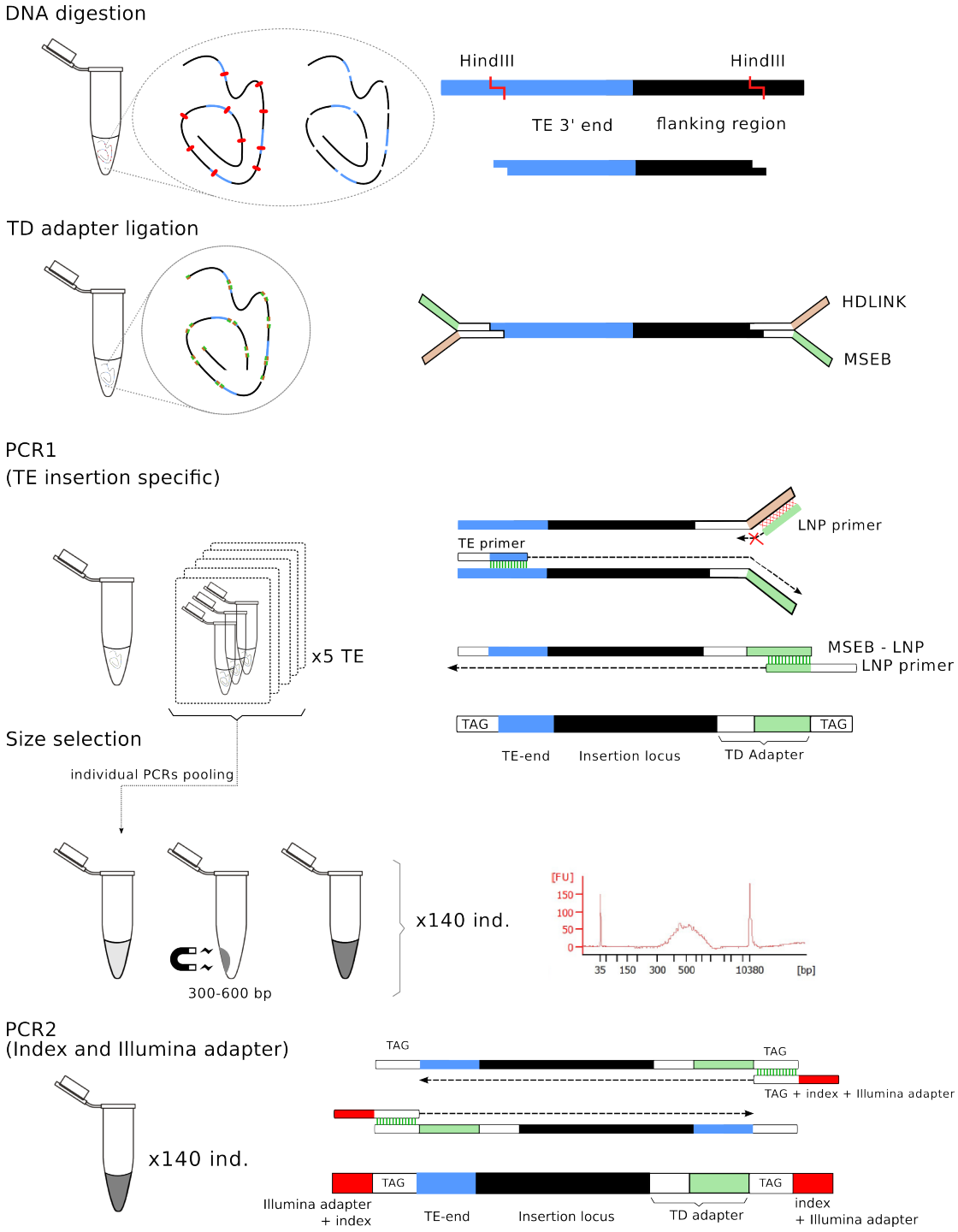



Figure 2 – Library preparation for high throughput sequencing of TD products. Genomic DNA was sheared using *HindIII*. Blue parts are copies of TE interspersed in the genome. Y shaped TD adapters were then ligated to cohesive ends. For each TE family and for each individual, PCR 1 was performed using a TE specific primer annealing to the end of each TE copy; subsequent elongation completed the complementary 3' end of the MSEB part of the adapter, allowing annealing of the LNP primer. For each TE family, three independent PCR were performed, and all PCR 1 products from one individual were pooled in an unique sample. Size selection using magnetic bead was then done fore each individual before normalization. Finally, PCR2 was performed for each individual pool in order to ligate indexes and Illumina adapters

paired-end sequenced on an Illumina Hiseq 2000 (1 lane) at the GeT-PlaGe core facility (Genome and Transcriptome, Toulouse) using TruSeq PE Cluster Kit v3 (2x100 bp) and TruSeq SBS Kit v3.

**Bioinformatic treatment of TD sequencing** The different steps of the informatics treatment from the raw sequencing dataset to population binary matrices for presence/absence of TE insertions per individual are described in Figure 3. A total number of 102,319,300 paired-end 101bp Illumina reads were produced by sequencing of PCR products. First, the reads pairs of each individual were quality checked and trimmed using `UrQt` v. 1.0.17 (Modolo et al. 2015) using standard parameters and a t quality threshold of 10. Reads pairs were then checked and trimmed for Illumina adapter contamination using `cutadapt` (Martin 2011). Specific amplification of TE insertion was controlled checking for expected TE end sequence on the R1 read pair using `Blat` (Kent 2002) with an identity threshold of 0.90. Only reads with an alignment-length/read-length ratio $\geq$ 0.90 were then retained. R2 reads for which the R1 mate passed this filter were then selected for the insertion loci construction, after the removal of the TD adapter on the 5' start using `cutadapt` and the removal of reads under 30 bp. Selected reads were separated in each individual according to the TE families for loci construction.

In order to correct the inter-individual coverage variations, we performed a sampling of the cleaned reads as follow. First, for each TE family, distribution of the number of read per individual was drown, and individuals with less reads reads than the first decile of this distribution were removed; then cleaned reads of the remaining individuals were randomly sampled at the value of the first decile of coverage (this value varies among TEs).

For each TE, the sampled reads of each retained individual were clustered together using the `CD-HIT-EST` program (Li et Godzik, 2006) to recover insertion loci. During this all-to-all reads comparison, the alignments must had a minimum of 90 percent identity, and the shortest sequence should be 95% length of the longest, global identity was used and each read was assigned at its best cluster (instead of the first that meet the threshold). In a second step, the reference reads of each locus within individual, given by `CD-HIT-EST`, were clustered with all the references reads of all individuals, using the same threshold, in order to build the locus catalog including list of loci of all individuals and the coverage for each locus in each individual. After this step, insertion loci that matches known repeats from Goubert et al. (2015) were discarded; alignments were performed with `Blastn` (Altschul et al. 1990) using default parameters.

Since the quality control removed a substantial number of reads for the construction of TE insertions catalog, the raw R2 reads (TD adapter removed), that could have been discarded in a first attempt were then mapped over the catalog in order to increase the scoring sensibility. Before mapping, the raw R2 reads were also sampled at the first decile of individual coverage (see upper). At this step, individuals that have been previously removed from at least two TE families for loci construction were definitively removed from the whole analysis. Mapping was performed over all insertion loci of all TE families in a single run in order to prevent multiple hits. `Blat` (Kent, 2002) was used with an identity threshold of 90 percent. Visual inspection of alignment quality over 30 sampled loci per TE family was performed in order to ensure the quality of scoring.

Finally, for each TE family, a scored locus had to be shared by at least 2.5% of individuals and its coverage as well as the variance of this coverage had to be below the 99th centile of their distribution among loci (excluding the "0"s). After those filters, remaining loci were turned to binary scoring (1 = presence, 0 = absence) for genetic analyses.

In order to check if the sampling procedure would affect our results, the read sampling procedures and subsequent analysis were performed independently 3 times.

**Genetic analyses and Genomic scan.** Population structure analyses were performed independently for each TE. Principal Coordinate Analysis (PCoAs) were performed to identify genetic clusters using the `ade4` package (Dray et Dufour, 2007) of R vers. 3.2.1 (R development core team 2015). S7 coefficient of Gower and Legendre was used as genetic distance since it gives more weight to shared insertions and shared absences were not used, since they do not give information about the genetic distance between individuals. Pairwise populations $F_{ST}$ were computed using `Arlequin 3.5` (Excoffier et Lischer, 2010); significance of the index was assessed over 1000 permutations using a significance threshold of

0.05.

The genomic scan was performed in two steps for each of the replicates of each TE. First, `Bayescan 2.1` (Foll et Gaggiotti, 2008) was used to test for each locus deviation from neutrality. Bayescan consider a fission/island model where all subpopulations derive from a unique ancestral population. In this model, variance in allele frequencies between subpopulations is expected to be due either to the genetic drift that occurred independently in each subpopulation or to selection that is a locus-specific parameter. The differentiation at each locus in each subpopulation from the ancestral population is thus decomposed into a $\beta$ component (shared by all loci in a subpopulation) and is related to genetic drift, and an $\alpha$ component (shared for a locus by all subpopulations) due to selection. Using a Bayesian framework, `Bayescan` tests for each locus the significance of the $\alpha$ component. Rejection of the neutral model a one locus is done using posterior Bayesian probabilities and controlled for multiple testing using false discovery rate. In addition, `Bayescan` manage uncertainty about allele frequency from dominant data such as the TD polymorphism, leaving freely vary the $F_{IS}$ during the estimation of parameters. `Bayescan` was used with default values except for the prior odds that were set to 100 (more compatible with datasets with a large number of loci, see `Bayescan` manual), and a significance $q$-value threshold of 0.05 was used to retain outliers loci. In a second step, only outliers loci suggesting divergent directional selection between, Europe and Vietnam were considered. To identify them, locus by locus Analyses of Molecular Variance (AMOVAs) were performed using `Arlequin 3.5` for each TE family. Significance of the $F_{CT}$ (inter group differentiation) between Vietnamese and European populations was assessed performing 10,000 permutations with a significance threshold of 0.05. For each dataset, `Bayescan` outliers were crossed with significant $F_{CT}$ loci to retain candidate loci.

**Outlier analyses and PCR validation.** Attempts to identify the genomic environment of the candidate loci were performed by mapping the outlier sequences (reference R2 read) onto the assembled consensus transcriptome of *Ae. albopictus* (Armbrouster and Poelchau, available at http://www.albopictusexpression.org/) and the assembled genome of *Aedes aegypti* (Nene *et al.*, 2007) using `Blast`. `Blastn` alignments were performed with default parameters and sorted according to alignment score.

Pairs of primers were designed for each outlier locus in order to be used in standardized conditions. Forward primer was located in the TE end of the concerned family and reverse primer was set from the outlier locus (pairs of primers for successfully amplified insertions are provided in supplementary data). Primer pairs were first tested on a set of 10 individuals in order to assess their specificity using 1/50 dilution of starting DNA from the TD experiment. Validated primers were then used to check the insertions polymorphism in 47 representatives individuals from the 8 populations studied in the TD experiment using 1/50 dilutions of the starting DNA (not all individuals could be used because of DNA limitations). All PCRs were conducted in a final volume of 25$\mu$L using 0.5$\mu$L of diluted DNA, 0.5$\mu$L of each primer (10$\mu$M), 1$\mu$l of dNTPs (10mM) and 1 U of DreamTaq Polymerase with 1X green buffer (ThermoFisher Scientific). Amplification was performed as follows: denaturation at 94 °C for 2 minutes then 34 cycles including denaturation at 94 °C for 30 seconds, hybridization at 60 °C for 45 seconds and elongation at 72 °C for 45 seconds; final elongation was performed for 10 minutes at 72 °C. After 45 minutes migration of the PCR product on 1X electrophoresis agarose gel, insertion polymorphism was assessed independently by CG and MB.

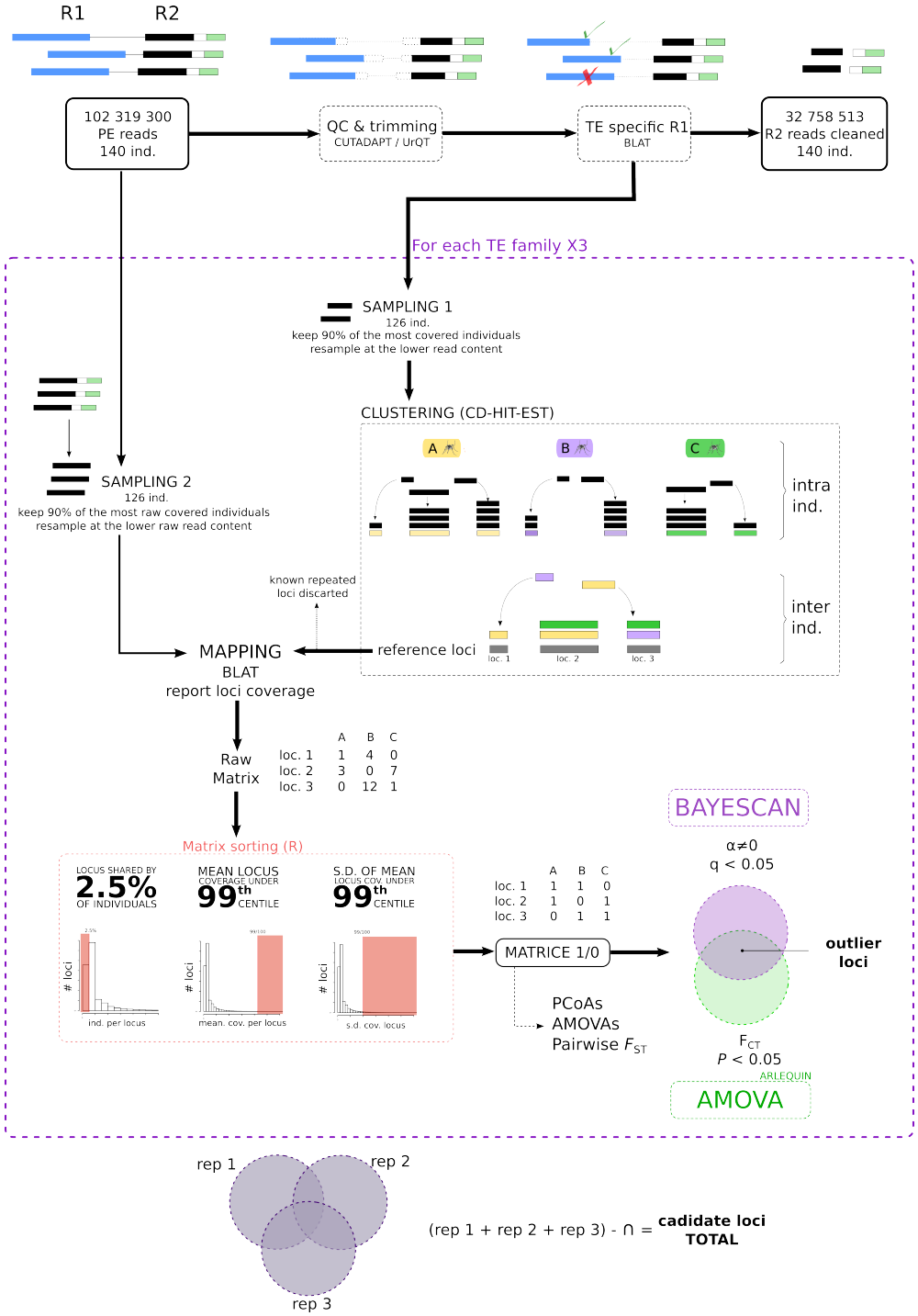



Figure 3 – Bioinformatic workflow from sequencing to outlier analysis. Details are given in the main text. R1 reads include the terminal end of one TE copy (in blue) and R2 reads include flancking region (black) and the TD adapter (white and green). After cleaning, R2 flanking regions are sampled to account for the coverage heterogeneity among individuals (sampling 1). Individual clustering of these reads, reference sequences (colored) of each insertion loci are then clustered between individuals to build the insertion catalog. Raw R2 reads are then sampled and mapped over the catalogs (1 catalog per TE family) and individual coverage per locus is calculated. Locus are then filterd for insertion frequency and coverage before population genetics and outlier analysis on 1/0 matrices. Steps surrounded in puprle dashed line are replicated three times. At the end, all outliers from all replicates are recovered (candidate loci total).

Table 3 – Total number of loci per TE family for each of the three sampling replicate (M1, M2, M3)

| TE | IL1 | L2B | RTE5 | RTE4 | Lian1 | Total |
|---|---|---|---|---|---|---|
| M1 | 11026 | 24290 | 40388 | 20975 | 31696 | **128375** |
| M2 | 10942 | 24150 | 40905 | 20918 | 31566 | **128481** |
| M3 | 11046 | 24241 | 40903 | 20805 | 31622 | **128617** |

# Results

## High throughput TD genotyping

Illumina sequencing of all TD amplification products produced a total amount of 102,319,300 paired-end reads (2x100bp). After quality and specificity filtering 24,332,715 R2 reads were suitable for sampling and insertion loci clustering. On average, a total number of 128,491 polymorphic insertion loci were available for each of the 3 replicates. Details for each TE and each replicates are given in table 3.

The mean number of loci per individual and per TE ranged from 1025±290 s.d. (IL1, mean and s.d. averaged over the 3 replicates) to 3266±766 s.d. (RTE5). Details are given in table 4.

## Population structure

**PCoAs.** Principal Coordinates Analyses were performed independently for each of the 5 TEs (Figure 4). Among the 3 main Principal Coordinates (PCs), individuals tends to be grouped according to their respective populations with little overlaps between groups. However, the mains PCs represents only a small fraction ($< 10\%$) of total genetic variation, suggesting a weak genetic structuring between the populations. Overall, individuals from Vietnamese populations (HCM, TA, VT) are grouped together in a single cluster, at the remarkable exception of 13 to 14 individuals from HCM that tends to be more related to European individuals at the L2B and RTE5 markers than with the three others. BCN individuals (Spain) represents the most homogeneous/aggregated group, that is both well differentiated from Vietnamese and French individuals (SP, CGN, NCE and PLV).

**AMOVA.** In agreement with PCoAs analysis, AMOVAs attributed very few genetic variance among groups (Vietnam-Europe) and between populations within groups (Table 5). In the studied populations, most of the genetic variance was located among individuals within groups.

**Pairwise FST.** Measures of genetic differentiation among pairs of populations were consistent with previous results (Table 6): BCN population shows the highest FST values with the other populations, for each of the five TEs, while Vietnamese populations were the most closely related. While VT is located 100km away from TA and HCM (both sampled in the same city, Hô Chi Minh) the FST values are very similar between the three Vietnamese populations, suggesting no influence geography at this scale. CGN and NCE, that were sampled in the same urban area (Nice agglomeration) are also little or even not significantly differentiated depending on the TE family. The previously identified intermediate pattern of HCM with some European populations at L2B and RTE5 loci (PCoAs analyses) is also found at the FST level, especially regarding the low differentiation with the PLV population for these markers.

## Genomic scan

Research of outlier insertion polymorphisms for both selection signature in the Bayescan model (island model) and for significant $F_{CT}$ (using `Arlequin`) between Vietnam-Europe groups identified 92 candidate insertion loci (Figures 5 and 6). Most of the insertions are found in both area (no private allele), except for RTE4_6 and RTE4_7 that were not found in Vietnam (Figure 6). In addition, a majority of outliers corresponds to high frequency insertions in Europe, while the same trend is not observed at 92 randomly chosen loci among those having the same minimum insertion frequency ($\geq$ 20 individuals/locus) between Europe and Vietnam (Figure 7).

PCR amplification was successful at the first attempt for 12 outlier loci, for which the insertion pattern in 47 representative individuals confirmed the outlier pattern obtained from TD (see supplementary material).

Table 4 – Mean number of loci per individual (loci/ind.) with standard deviation (s.d.) and mean number of individuals that share a locus (ind./loc.) with s.d. for each TE family and each of the three sampling replicate (M1, M2, M3). Mean and s.d. values are averaged over the the three replicates (bold)

| TE | IL1 | | L2B | | RTE5 | | RTE4 | | Lian1 | |
|---|---|---|---|---|---|---|---|---|---|---|
| *mean* | loci/ind. | ind./loc. | loci/ind. | ind./loc. | loci/ind. | ind./loc. | loci/ind. | ind./loc/ | loci/ind. | ind./loc. |
| M1 | 1025.86 | 11.04 | 2396.27 | 11.85 | 3263.17 | 9.59 | 2227.34 | 12.65 | 2588.55 | 9.66 |
| M2 | 1008.24 | 10.97 | 2407.17 | 11.96 | 3260.34 | 9.56 | 2201.50 | 12.52 | 2567.63 | 9.68 |
| M3 | 1041.06 | 11.22 | 2424.96 | 12.00 | 3273.78 | 9.60 | 2212.50 | 12.66 | 2574.26 | 9.69 |
| **average** | **1025.05** | **11.08** | **2409.46** | **11.94** | **3265.76** | **9.58** | **2213.78** | **12.61** | **2576.81** | **9.68** |
| *s.d.* | | | | | | | | | | |
| M1 | 290.86 | 15.83 | 484.91 | 16.82 | 765.14 | 12.85 | 732.95 | 16.02 | 840.26 | 12.92 |
| M2 | 286.28 | 15.67 | 489.70 | 16.97 | 766.65 | 12.74 | 728.84 | 15.84 | 830.93 | 13.11 |
| M3 | 292.02 | 16.15 | 495.43 | 16.96 | 765.85 | 12.93 | 733.69 | 15.94 | 833.92 | 13.00 |
| **average** | **289.72** | **15.88** | **490.01** | **16.92** | **765.88** | **12.84** | **731.83** | **15.93** | **835.04** | **13.01** |

Table 5 – Analyses of Molecular Variance (AMOVAs) for the three replicates (M1,M2,M3) of read sampling for the five TE families (IL1, L2B, RTE5, RTE4, Lian1). Values are given in percentage of the total genetic variance.

| | IL1 | | | L2B | | | RTE5 | | | RTE4 | | | Lian1 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | M1 | M2 | M3 | M1 | M2 | M3 | M1 | M2 | M3 | M1 | M2 | M3 | M1 | M2 | M3 |
| Among groups | 0.70 | 0.67 | 0.59 | 1.29 | 1.26 | 1.22 | 1.09 | 1.08 | 1.10 | 1.97 | 2.03 | 2.04 | 0.74 | 0.67 | 0.67 |
| Among populations within groups | 5.15 | 5.27 | 5.37 | 3.63 | 3.58 | 3.60 | 3.40 | 3.37 | 3.36 | 6.74 | 6.67 | 6.78 | 4.47 | 4.56 | 4.52 |
| Within populations | 94.16 | 94.06 | 94.04 | 95.08 | 95.16 | 95.18 | 95.51 | 95.55 | 95.54 | 91.29 | 91.30 | 91.18 | 94.79 | 94.77 | 94.81 |

Table 6 – Pairwise $F_{ST}$ estimates between European (BCN, CGN, NCE, PLV, SP) and Vietnamese (HCM, TA, VT) populations of *Ae. albopictus* for each of the 5 TE family. Values are computed from the first (M1) replicate. Italicized values mean that at least one of the three replicate has a non significant estimate for the comparison ($P \leq 0.05$, `Arlequin 3.5`, 1000 permutations)

| **IL1** | BCN | CGN | NCE | PLV | SP | HCM | TA |
|---|---|---|---|---|---|---|---|
| CGN | 0.098 | | | | | | |
| NCE | 0.109 | 0.015 | | | | | |
| PLV | 0.079 | 0.026 | 0.036 | | | | |
| SP | 0.128 | 0.040 | 0.027 | 0.073 | | | |
| HCM | 0.098 | 0.024 | 0.027 | 0.053 | 0.053 | | |
| TA | 0.075 | 0.042 | 0.037 | 0.055 | 0.067 | 0.030 | |
| VT | 0.088 | 0.025 | 0.027 | 0.047 | 0.052 | 0.013 | 0.016 |

| **L2B** | BCN | CGN | NCE | PLV | SP | HCM | TA |
|---|---|---|---|---|---|---|---|
| CGN | 0.063 | | | | | | |
| NCE | 0.060 | 0.024 | | | | | |
| PLV | 0.058 | 0.015 | 0.029 | | | | |
| SP | 0.087 | 0.015 | 0.031 | 0.049 | | | |
| HCM | 0.061 | 0.023 | 0.033 | 0.020 | 0.046 | | |
| TA | 0.060 | 0.046 | 0.038 | 0.049 | 0.060 | 0.023 | |
| VT | 0.073 | 0.039 | 0.041 | 0.050 | 0.054 | 0.012 | 0.021 |

| **RTE5** | BCN | CGN | NCE | PLV | SP | HCM | TA |
|---|---|---|---|---|---|---|---|
| CGN | 0.069 | | | | | | |
| NCE | 0.055 | 0.028 | | | | | |
| PLV | 0.052 | 0.011 | 0.023 | | | | |
| SP | 0.090 | 0.020 | 0.033 | 0.049 | | | |
| HCM | 0.051 | 0.022 | 0.025 | 0.013 | 0.049 | | |
| TA | 0.056 | 0.043 | 0.030 | 0.039 | 0.060 | 0.017 | |
| VT | 0.066 | 0.041 | 0.038 | 0.040 | 0.059 | 0.012 | 0.015 |

| **RTE4** | BCN | CGN | NCE | PLV | SP | HCM | TA |
|---|---|---|---|---|---|---|---|
| CGN | 0.096 | | | | | | |
| NCE | 0.103 | *0.012* | | | | | |
| PLV | 0.108 | 0.051 | 0.063 | | | | |
| SP | 0.148 | 0.060 | 0.054 | 0.102 | | | |
| HCM | 0.114 | 0.047 | 0.049 | 0.086 | 0.082 | | |
| TA | 0.110 | 0.040 | 0.043 | 0.088 | 0.097 | 0.030 | |
| VT | 0.119 | 0.047 | 0.044 | 0.086 | 0.066 | 0.021 | 0.029 |

| **Lian1** | BCN | CGN | NCE | PLV | SP | HCM | TA |
|---|---|---|---|---|---|---|---|
| CGN | 0.068 | | | | | | |
| NCE | 0.075 | *0.009* | | | | | |
| PLV | 0.057 | 0.024 | 0.027 | | | | |
| SP | 0.109 | 0.034 | *0.026* | 0.065 | | | |
| HCM | 0.064 | 0.018 | 0.021 | 0.045 | 0.051 | | |
| TA | 0.057 | 0.036 | 0.028 | 0.049 | 0.067 | 0.017 | |
| VT | 0.077 | 0.034 | 0.034 | 0.054 | 0.055 | 0.022 | 0.025 |

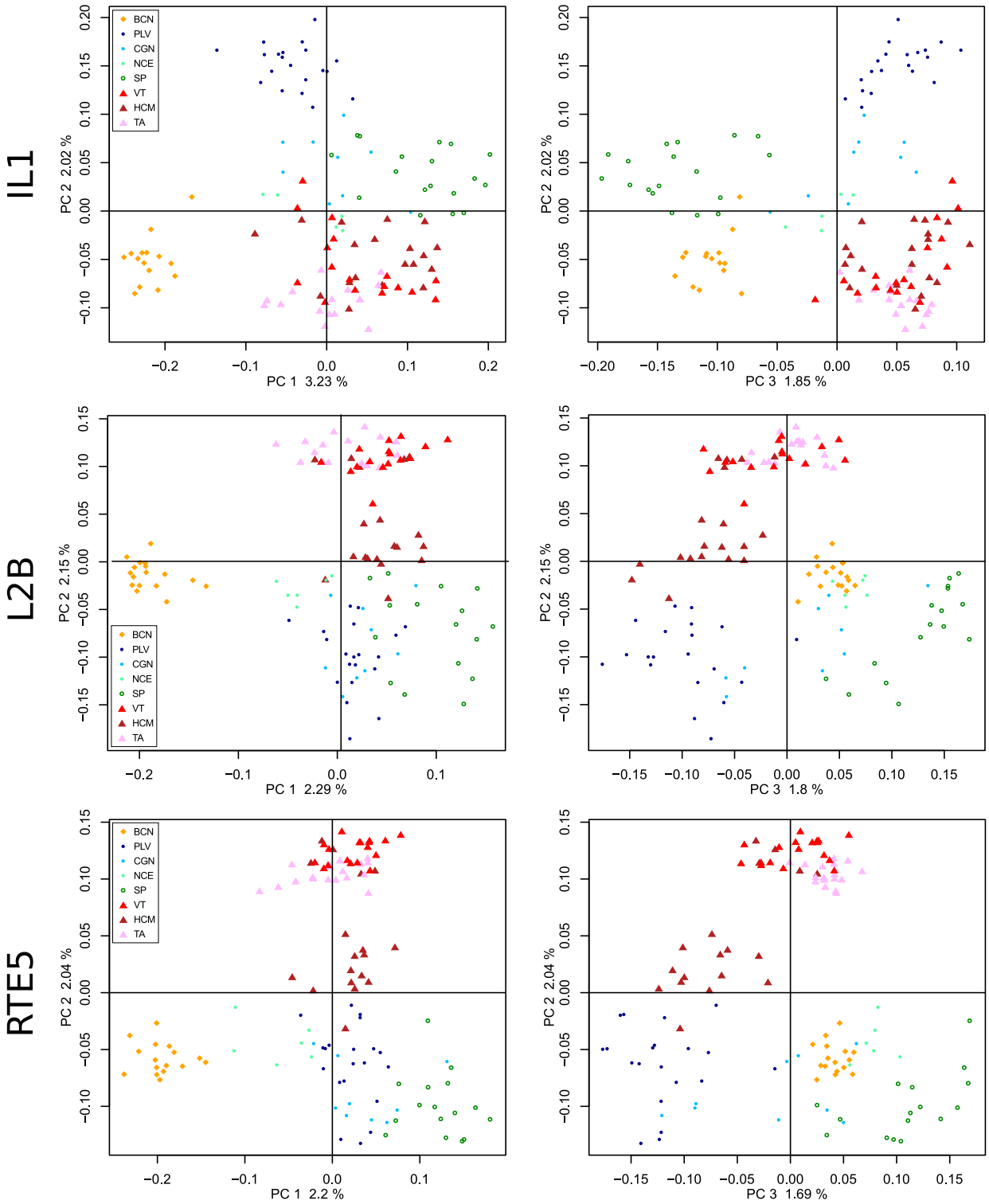



Figure 4 – Projection of individuals over the three first principal coordinates (PC) of Principal Coordinates Analyses (PCoAs) computed over all loci for each of the TE family (replicate M1). Proportion of inertia represented by each axe is noted in %. ◇ ○ ● European populations; △ Vietnamese populations

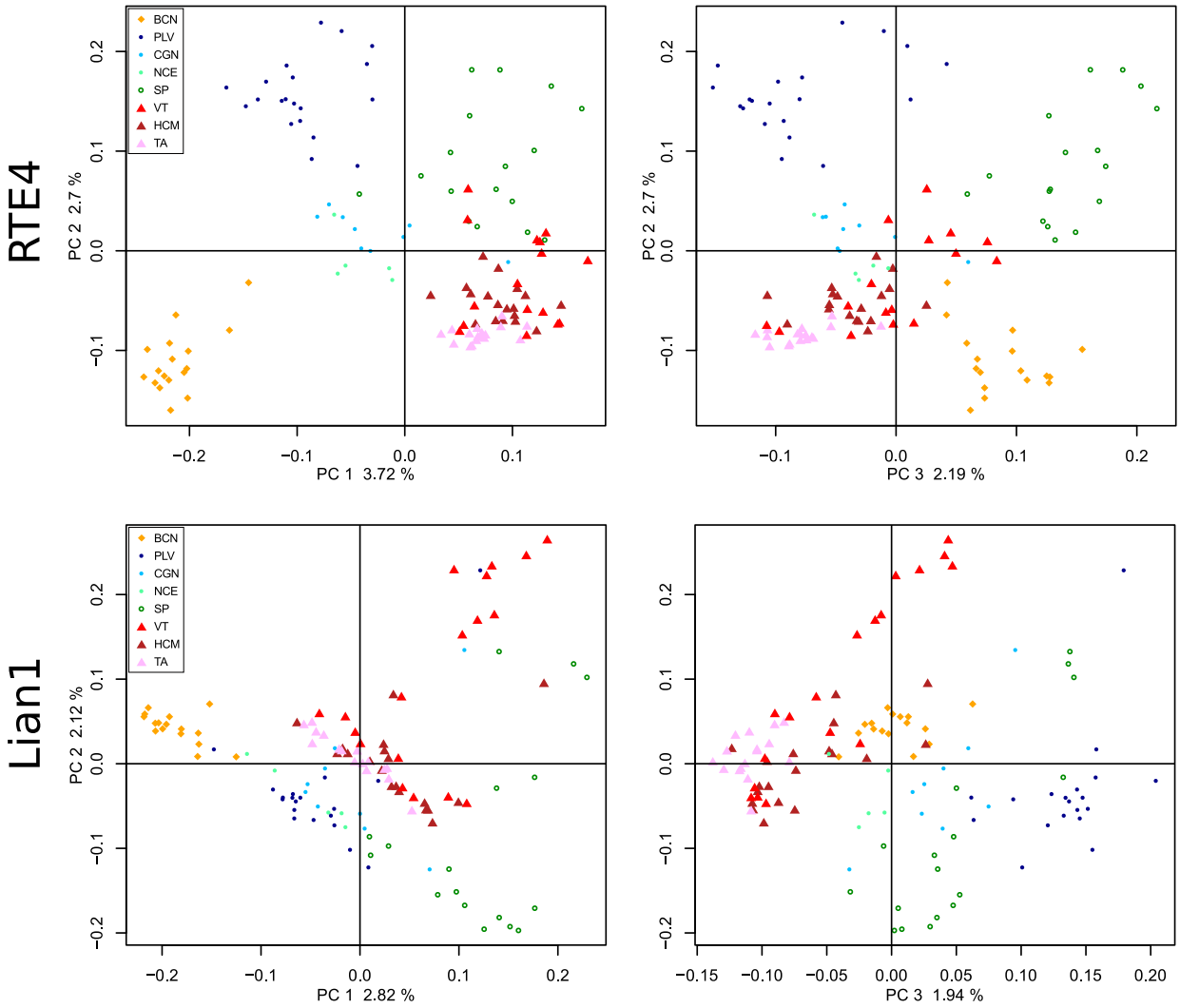



Figure 4 – Continued – Projection of individuals over the three first principal coordinates (PC) of Principal Coordinates Analyses (PCoAs) computed over all loci for each of the TE family (replicate M1). Proportion of inertia represented by each axe is noted in %. ◇ ○ ● European populations; △ Vietnamese populations

## Discussion

The goal of our study was to identify genomic regions involved in adaptive evolution of *Ae. albopictus* thanks to the development of new genetic markers. From the high-throughput genotyping of 5 TE families insertion polymorphisms, we identified up to 128,617 polymorphic loci among a hundred of individuals from 8 sampling sites. While some genome size variations has been suggested in this species, we can confidently consider that the average genome size of *Ae. albopictus* exceeds one billion of base-pairs (Rao et Rai, 1987; Kumar et Rai, 1990; Goubert *et al.*, 2015; Dritsou *et al.*, 2015). Accordingly, the exceptional amount of markers scored in this study offer a comfortable genomic density of one marker every 10 kb. In addition, the TEs used for this study are non-LTR (LINE) retrotransposons that usually do not show a specific insertion site preference (Malik *et al.*, 1999) making them likely to be well dispersed in the genome.

We provide here a new and cost efficient method to quickly generate a large amount of polymorphic markers without extensive knowledge about one species genome. Specifically, this method could be extremely valuable for species with a large genome size, where TE density could severely compromise the development of more classical approaches, such as the very popular RAD-sequencing (Miller *et al.*, 2007).

The genetic structure of the studied populations showed strong consistency between sampling replicates of individuals' reads, demonstrating the robustness of the method in spite of an initial substantial coverage variation among individuals. Population genetics analyses revealed a very low level of genetic structuring between European and Vietnamese populations. Among the studied populations, AMOVAs showed that most of the genetic variation is distributed between individuals within populations ($>$ 90%), and as suggested by pairwise $F_{ST}$ and PCoAs, only a small part ($<$ 10%) of the genetic variance is due to differentiation between populations. This singular population structure is in agreement with previous results gathered in *Ae. albopictus* using different collection of allozymes, mtDNA or microsatellites markers (Black *et al.*, 1988b,a; Kambhampati *et al.*, 1991; Zhong *et al.*, 2013; Gupta et Preet, 2014; Manni *et al.*, 2015).

Moreover, a recent analysis performed with a set of 10 microsatellites on individuals from the same populations (at the exception of BCN) showed a similar distribution of genetic variation among hierarchical levels (Minard *et al.*, 2015). In this study, the $F_{CT}$ value (hierarchical analogue of $F_{ST}$ between group of populations) between French and Vietnamese populations was 2.61%. Depending on the TE family, our values ranges from 0.67 to 2%; the observed differences between our study and microsatellites data could be attributed to several factors such as the addition of the divergent BCN population, the introduction of false negative scoring in our data due to the correction of the original heterogeneity in the sequencing coverage or to a more accurate estimation of genetic variation thanks to a much larger amount of loci. However, these results demonstrate the reliability of our markers and confirm that the global genetic structure most likely fits a non-hierarchical island model. The genetic differentiation we measured is indeed as high among European populations as between populations from Europe and Vietnam. The BCN sample appears to be the most differentiated one. Those mosquitoes were sampled in Sant Cugat del Vallès, Spain, were the first invasive population of *Ae. albopictus* has been recorded in this country in 2004 (Aranda *et al.*, 2006). This place is located in the neighborhood of Barcelona, an important commercial port in the Mediterranean sea where it was probably introduced. In France, the first populations settled near Nice in 2004 (Delaunay *et al.*, 2009), where our populations NCE and CGN are located. These places also have intensive port activities and are located close to northern Italy, where *Ae. albopictus* is present at high density since the 1990's (Medlock *et al.*, 2012). The genetic diversity could be compatible with a scenario of multiple and independent introductions, as suggested *Ae. albopictus* (Urbanelli *et al.*, 2000; Birungi et Munstermann, 2002; Takumi *et al.*, 2009; Becker *et al.*, 2013). This pattern could however be also the result of founder events that could occur during colonisation and/or a restriction of the genes flow between the populations consecutive to their introduction. Answering such a question would require an extended sampling among the native area.

The three Vietnamese populations appear to be more homogeneous than the European ones, at the notable exceptions of individuals from the VT population with the IL1 TE and most of the individuals from the HCM sampling site for the TE L2B and RTE5 that seems to be closer to the European popu-

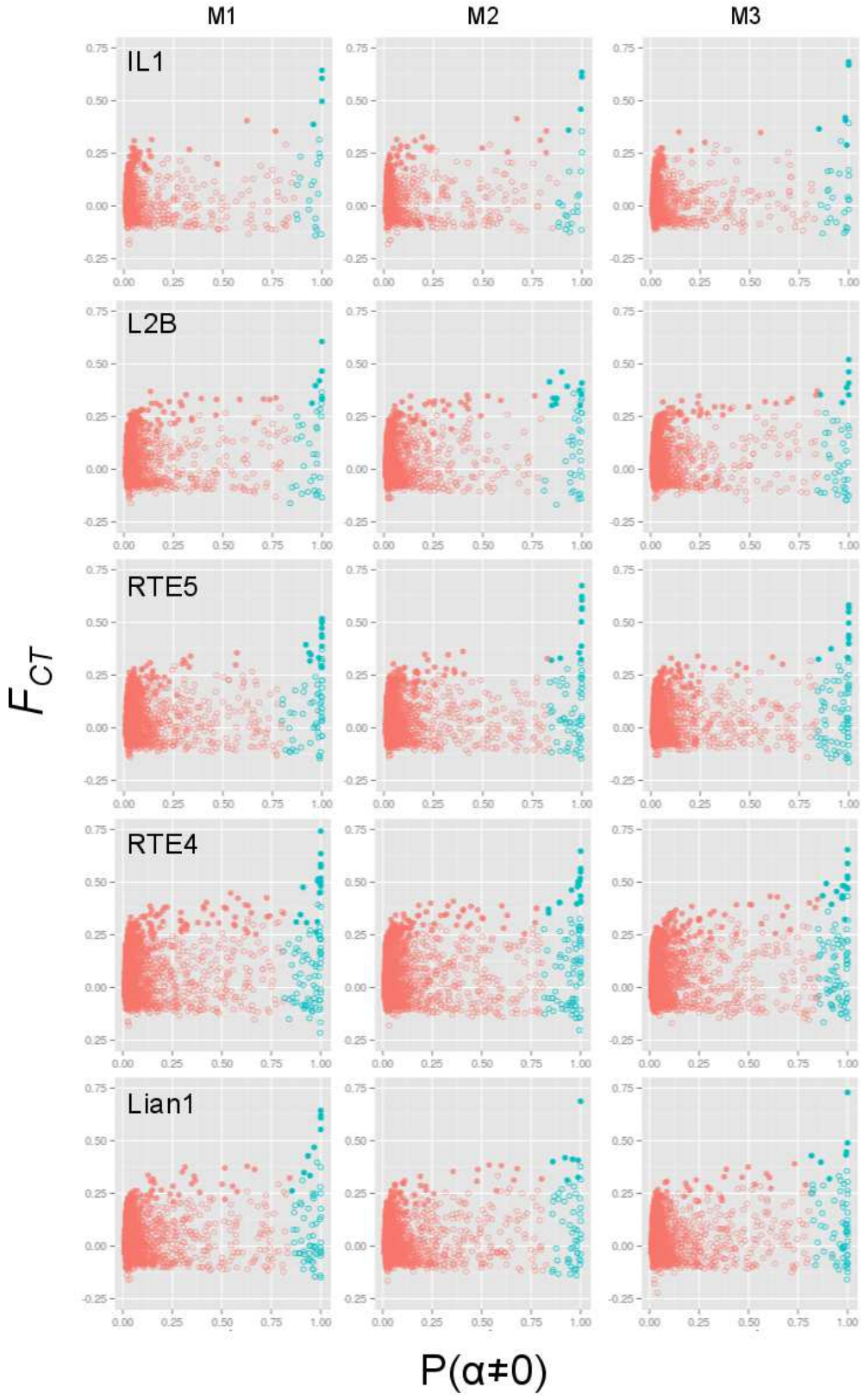



Figure 5 – Locus by locus $F_{CT}$ according to posterior Bayescan probability of being under selection ($\alpha \neq 0$) for each TE family (IL1, L2B, RTE5, RTE4, Lian1) in row and each replicate of initial read sampling (M1, M2, M3) in column. Significant FCT are labelized by a plain • ; significant $P(\alpha \neq 0)$ are in green.

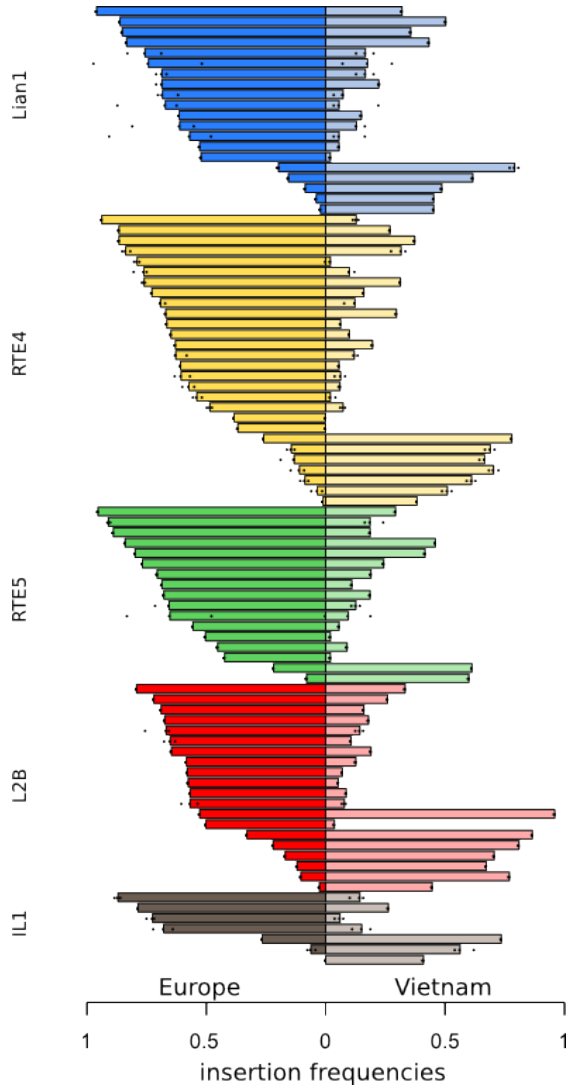



Figure 6 – Insertion frequencies in Europe and Vietnam for the 92 outliers loci. Bars represent the median value from the three reads sampling replicates and dots the values from the other replicates (if found). Colors correspond to each of the 5 TE family

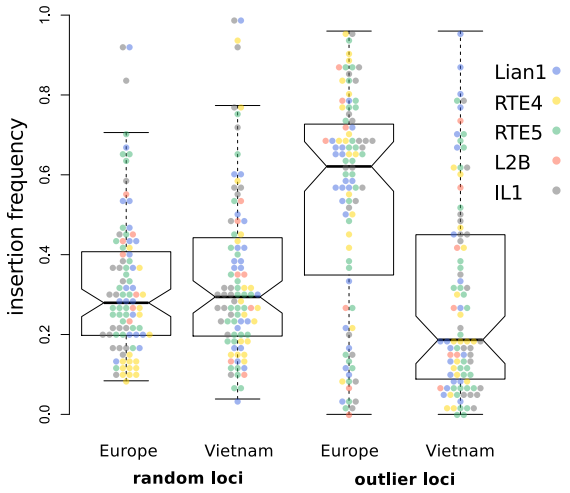



Figure 7 – Insertion frequencies of 92 randomly chosen loci among those having the same minimum insertion frequency (≥20 individuals) as outliers compared to the 92 outlier loci. Random loci were taken from the first replicate (M1) and values for outliers are median values obtained among the three replicates. Non-overlapping notches indicate a significant difference between the true medians (thick dark horizontal bars)

lations. Such an admixture pattern could be caused by a past or ongoing gene flow between European and HCM populations. Indeed, the differences observed between the TE families could be due to different dynamics of transposition, and maybe reflect the population structure at different times. Accordingly, we never mixed markers from different TE families in our analyses.

Outlier analysis revealed 92 loci with high posterior probabilities of being under positive selection between European and Vietnamese populations. When possible, the PCR amplification of the outlier loci using a set of representative individuals always confirmed a shift of insertion frequencies toward either the European or the Vietnamese sampling sites. This suggest that in spite of a reduced coverage, introduced by sampling in the dataset, the scored insertion polymorphisms are reliable. In addition, our method is likely to be conservative: the `Bayescan` outliers were selected for their consistency with a significant $F_{CT}$ between European temperate and Vietnamese tropical populations, which avoid retaining outliers that we were not looking for, for example those due to a population specific event.

The omics resources currently available did not allowed an exhaustive characterization of the genetic environment at the neighborhood of the outlier insertions. Neither blast of the outlier loci over the assembled transciptome of *Ae. albopictus* (Armbrouster and Poelchau, available at http://www.albopictusexpression.org/) nor the assembled genome of *Ae. aegypti* (the closest species with an assembled genome, Nene *et al.*, 2007) allowed us to attribute outliers to a unique location. These results are not surprising: first, our main hypothe-

sis assume that most of the insertions occurred at a distance on actual target of selection and could thus locate in none-transcribed regions. Second, the genome of the relative *Ae. aegypti* appears to be far distant from *Ae. albopictus* while those species belongs to the same genus: when compared by blast alignments, the divergence of orthologous genes is less than 79% identity (Table 5 of Dritsou *et al.*, 2015).

Interestingly, we found significantly more outlier loci with a high frequency of TE insertion in Europe and low frequency in Vietnam than the opposite pattern. This was unexpected regarding our initial assumptions: a favored allele selected in one or another environment has no reason to be more often associated with the presence or the absence of a TE insertion at linked sites. However, we found that the vast majority of the sequenced TE insertions (whole dataset) segregate at low frequencies (around 10% of all individuals). While our reads sampling procedure could have artificially lowered the mean insertion frequency of the loci, this effect should be small because in our final datasets the TE insertions frequencies (= the number of individuals that share an insertion) are not correlated with the mean number of read per individual at the considered locus (see supplementary material). When considering the linked region of one polymorphic TE insertion, if a favorable mutation appears in an individual where the insertion is absent, the increase of frequency of this "absence" haplotype will thus most of the time have a modest effect on the genetic differentiation at this marker, since it is already segregating at high frequency. By contrast, if a favorable mutation appears in a TE "presence" haplotype, the increase in frequency of the linked TE insertion would lead to high $F_{ST}$ ($F_{CT}$) values. In absence of an alternative explanation, our outlier loci could thus indicate in which subset of populations the adaptive mutation occurred, and in the present case, this would have happened more frequently in the temperate populations.

Two scenarios, not mutually exclusive, could be invoked in the light of our data. A simple case would be a direct adaptive evolution in European invasive population that originated from tropical regions of the native area. A second hypothesis, could be that invasive temperate populations came from northern-most territories of the native area such as northern China or Japan (Hawley *et al.*, 1987; Kambhampati *et al.*, 1991; Urbanelli *et al.*, 2000; Zhong *et al.*, 2013). In both cases, the temperate populations would be derived ones. *Ae. albopictus* is supposed to originate from tropical forests, where its closest relatives of the Albopicta group are currently found (Hawley, 1988). In a recent study, Porretta *et al.* (2012) suggested using new variable COI mtDNA sequences and historical species range modeling that northern areas of the native range of *Ae. albopictus* would be the latest to have been colonized after a range expansion from refugees following the last glacial maximum (≈21,000 years ago). The authors suggested that *Ae. albopictus* may have followed the human populations during their expansion from south to north in this area, that began ≈15,000 years ago. Thus wherever the origin of the invasive individuals sampled in Europe, it is likely that they are representatives of populations that had recently undergo a shift of selective pressure from tropical to temperate climatic conditions. A way to distinguish between these possibilities would be to search if the same outlier insertions are present in several temperate populations from the native area.

It is important to notice that the results presented here only lie for a subset of the Asian tiger mosquito populations located in temperate and tropical environments. It is thus probable that some of the outlier detected could be specific to this particular comparison and do not reflect the global pattern of differentiation between tropical and temperate populations. Research of the same candidate loci between other tropical and temperate populations from the native and non-native areas would be extremely valuable to extrapolate our results at a larger scale. Should the same outlier insertions be found at a high frequencies in temperate locations, such as in USA, Japan or China, extended investigations about the origin of invasive populations would help clarify if those similarities are due to a common origin of the individuals or cases of parallel evolution. This study already provide for some candidate loci a set of functional primers, that could be directly used to answer this question in any DNA sample of *Ae. albopictus*.

## Acknowledgements

We are grateful to Valèria Romero Soriano and her family for their help during sampling in Sant Cugat dèl Vallès. We also thanks Manon Vigneron for PCR validation experiments. This work was performed us-

ing the computing facilities of the CC LBBE/PRABI. C.G. received a grant from the French Ministry of Superior Education. This work was supported by the Centre National de la Recherche Scientifique, the Institut Universitaire de France, and preliminary experiments benefited from of grant of the Federation de Recherche 41 "Bio-Environnement et Santé". Original maps used to describe sampling where taken from d-maps.com at urls: `http://d-maps.com/carte.php?num_car=4719&lang=en` and `http://d-maps.com/carte.php?num_car=708&lang=en`.

## Conflict of interest

The authors declare no conflict of interest

## References

Aranda, C., Eritja, R., and Roiz, D. *First record and establishment of the mosquito Aedes albopictus in Spain.* **Medical and veterinary entomology**, 20(1):150–2, March 2006. doi: 10.1111/j.1365-2915.2006.00605.x.

Becker, N., Geier, M., Balczun, C., Bradersen, U., Huber, K., Kiel, E., Krüger, A., Lühken, R., Orendt, C., Plenge-Bönig, A., Rose, A., Schaub, G. A., and Tannich, E. *Repeated introduction of Aedes albopictus into Germany, July to October 2012.* **Parasitology research**, 112(4):1787–90, April 2013. doi: 10.1007/s00436-012-3230-1.

Biedler, J. and Tu, Z. *Non-LTR retrotransposons in the African malaria mosquito, Anopheles gambiae: unprecedented diversity and evidence of recent activity.* **Molecular biology and evolution**, 20(11):1811–1825, November 2003. doi: 10.1093/molbev/msg189.

Birungi, J. and Munstermann, L. E. *Genetic Structure of Aedes albopictus (Diptera: Culicidae) Populations Based on Mitochondrial ND5 Sequences: Evidence for an Independent Invasion into Brazil and United States.* **Annals of the Entomological Society of America**, 95(1):125–132, January 2002. doi: 10.1603/0013-8746(2002)095[0125:GSOAAD]2.0.CO;2.

Black, W. I. C. I. V., Hawley, W. A., Rai, K. S., and Craig, G. B. *Breeding structure of a colonizing species: Aedes albopictus (Skuse) in peninsular Malaysia and Borneo.* **Heredity**, 61(March):439–446, 1988a.

Black, W. C., Ferrari, J. A., and Sprengert, D. *Breeding structure of a colonising species : Aedes albopictus ( Skuse ) in the United States.* 60(April 1987), 1988b.

Bock, D. G., Caseys, C., Cousens, R. D., Hahn, M. A., Heredia, S. M., Hübner, S., Turner, K. G., Whitney, K. D., and Rieseberg, L. H. *What we still don't know about invasion genetics.* **Molecular Ecology**, pages n/a–n/a, December 2015. doi: 10.1111/mec.13032.

Bonin, A., Paris, M., Després, L., Tetreau, G., David, J.-P., and Kilian, A. *A MITE-based genotyping method to reveal hundreds of DNA polymorphisms in an animal genome after a few generations of artificial selection.* **BMC genomics**, 9:459, January 2008. doi: 10.1186/1471-2164-9-459.

Bonizzoni, M., Gasperi, G., Chen, X., and James, A. A. *The invasive mosquito species Aedes albopictus: current knowledge and future perspectives.* **Trends in parasitology**, 29 (9):460–468, September 2013. doi: 10.1016/j.pt.2013.07.003.

Boulesteix, M., Simard, F., Antonio-Nkondjio, C., Awono-Ambene, H. P., Fontenille, D., and Biémont, C. *Insertion polymorphism of transposable elements and population structure of Anopheles gambiae M and S molecular forms in Cameroon.* **Molecular ecology**, 16(2):441–452, January 2007. doi: 10.1111/j.1365-294X.2006.03150.x.

Casacuberta, E. and González, J. *The impact of transposable elements in environmental adaptation.* **Molecular ecology**, pages 1503–1517, January 2013. doi: 10.1111/mec.12170.

Colautti, R. I. and Lau, J. A. *Contemporary evolution during invasion: evidence for differentiation, natural selection, and local adaptation.* **Molecular Ecology**, 24(9):1999–2017, May 2015. doi: 10.1111/mec.13162.

Delaunay, P., Jeannin, C., Schaffner, F., and Marty, P. *News on the presence of the tiger mosquito Aedes albopictus in metropolitan France.* **Archives de pédiatrie : organe**

**officiel de la Sociéte française de pédiatrie**, 16 Suppl 2:S66–71, October 2009. doi: 10.1016/S0929-693X(09)75304-7.

Dray, S. and Dufour, A.-B. *The ade4 Package: Implementing the Duality Diagram for Ecologists.* **Journal of Statistical Software**, 22(4):1–20, September 2007. doi: 10.18637/jss.v022.i04.

Dritsou, V., Topalis, P., Windbichler, N., Simoni, A., Hall, A., Lawson, D., Hinsley, M., Hughes, D., Napolioni, V., Crucianelli, F., Deligianni, E., Gasperi, G., Gomulski, L. M., Savini, G., Manni, M., Scolari, F., Malacrida, A. R., Arcà, B., Ribeiro, J. M., Lombardo, F., Saccone, G., Salvemini, M., Moretti, R., Aprea, G., Calvitti, M., Picciolini, M., Papathanos, P. A., Spaccapelo, R., Favia, G., Crisanti, A., and Louis, C. *A draft genome sequence of an invasive mosquito: an Italian Aedes albopictus.* **Pathogens and global health**, page 2047773215Y0000000031, September 2015. doi: 10.1179/2047773215Y.0000000031.

Esnault, C., Boulesteix, M., Duchemin, J. B., Koffi, A. A., Chandre, F., Dabiré, R., Robert, V., Simard, F., Tripet, F., Donnelly, M. J., Fontenille, D., and Biémont, C. *High genetic differentiation between the M and S molecular forms of Anopheles gambiae in Africa.* **PloS one**, 3(4):e1968, January 2008. doi: 10.1371/journal.pone.0001968.

Excoffier, L. and Lischer, H. E. L. *Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows.* **Molecular ecology resources**, 10(3):564–7, May 2010. doi: 10.1111/j.1755-0998.2010.02847.x.

Foll, M. and Gaggiotti, O. *A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective.* **Genetics**, 180(2):977–93, October 2008. doi: 10.1534/genetics.108.092221.

Goubert, C., Modolo, L., Vieira, C., Valiente-Moro, C., Mavingui, P., and Boulesteix, M. *De novo assembly and annotation of the Asian tiger mosquito (Aedes albopictus) repeatome with dnaPipeTE from raw genomic reads and comparative analysis with the yellow fever mosquito (Aedes aegypti).* **Genome Biology and Evolution**, pages evv050–, March 2015. doi: 10.1093/gbe/evv050.

Gupta, S. and Preet, S. *Genetic differentiation of invasive Aedes albopictus by RAPD-PCR: Implications for effective vector control.* **Parasitology Research**, 113(6):2137–2142, 2014. doi: 10.1007/s00436-014-3864-2.

Handley, L.-J., Estoup, A., Evans, D. M., Thomas, C. E., Lombaert, E., Facon, B., Aebi, A., and Roy, H. E. *Ecological genetics of invasive alien species.* **BioControl**, 56(4):409–428, August 2011. doi: 10.1007/s10526-011-9386-2.

Hanson, S. M. and Craig, G. B. *Cold Acclimation, Diapause, and Geographic Origin Affect Cold Hardiness in Eggs of Aedes albopictus (Diptera: Culicidae).* **Journal of Medical Entomology**, 31(2):192–201, March 1994. doi: 10.1093/jmedent/31.2.192.

Hawley, W., Reiter, P., Copeland, R., Pumpuni, C., and Craig, G. *Aedes albopictus in North America: probable introduction in used tires from northern Asia.* **Science**, 236 (4805):1114–1116, May 1987. doi: 10.1126/science.3576225.

Hawley, W. A. *The biology of Aedes albopictus.* **Journal of the American Mosquito Control Association. Supplement**, 1:1–39, December 1988.

Kambhampati, S., Black, W. C., and Rai, K. S. *Geographic origin of the US and Brazilian Aedes albopictus inferred from allozyme analysis.* **Heredity**, 67 ( Pt 1)(September 1990):85–93, August 1991.

Kent, W. J. *BLAT–the BLAST-like alignment tool.* **Genome research**, 12(4):656–64, April 2002. doi: 10.1101/gr.229202.ArticlepublishedonlinebeforeMarch2002.

Kumar, A. and Rai, K. S. *Intraspecific variation in nuclear DNA content among world populations of a mosquito, Aedes albopictus (Skuse).* **TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik**, 79 (6):748–52, July 1990. doi: 10.1007/BF00224239.

Lande, R. *Evolution of phenotypic plasticity in colonizing species.* **Molecular ecology**, January 2015. doi: 10.1111/mec.13037.

Li, W. and Godzik, A. *Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences.* **Bioinformatics (Oxford, England)**, 22(13): 1658–9, July 2006. doi: 10.1093/bioinformatics/btl158.

Malik, H. S., Burke, W. D., and Eickbush, T. H. *The age and evolution of non-LTR retrotransposable elements.* **Molecular Biology and Evolution**, 16(6):793–805, June 1999. doi: 10.1093/oxfordjournals.molbev.a026164.

Manni, M., Gomulski, L. M., Aketarawong, N., Tait, G., Scolari, F., Somboon, P., Guglielmino, C. R., Malacrida, A. R., and Gasperi, G. *Molecular markers for analyses of intraspecific genetic diversity in the Asian Tiger mosquito, Aedes albopictus.* **Parasites & vectors**, 8(1):188, January 2015. doi: 10.1186/s13071-015-0794-5.

Medlock, J. M., Hansford, K. M., Schaffner, F., Versteirt, V., Hendrickx, G., Zeller, H., and Van Bortel, W. *A review of the invasive mosquitoes in Europe: ecology, public health risks, and control options.* **Vector borne and zoonotic diseases (Larchmont, N.Y.)**, 12(6):435–447, June 2012. doi: 10.1089/vbz.2011.0814.

Miller, M. R., Dunham, J. P., Amores, A., Cresko, W. A., and Johnson, E. A. *Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers.* **Genome research**, 17(2):240–8, February 2007. doi: 10.1101/gr.5681207.

Minard, G., Tran, F.-H., Tran-van, V., Goubert, C., Bellet, C., Lambert, G., Khanh, H. K. L., Huynh, T., Mavingui, P., and Valiente Moro, C. *French invasive Asian tiger mosquito populations harbor reduced bacterial microbiota and genetic diversity compared to Vietnamese autochthonous relatives.* **Frontiers in Microbiology**, 6, 2015. doi: 10.3389/fmicb.2015.00970.

Nene, V., Wortman, J. R., Lawson, D., Haas, B., Kodira, C., Tu, Z. J., Loftus, B., Xi, Z., Megy, K., Grabherr, M., Ren, Q., Zdobnov, E. M., Lobo, N. F., Campbell, K. S., Brown, S. E., Bonaldo, M. F., Zhu, J., Sinkins, S. P., Hogenkamp, D. G., Amedeo, P., Arensburger, P., Atkinson, P. W., Bidwell, S., Biedler, J., Birney, E., Bruggner, R. V., Costas, J., Coy, M. R., Crabtree, J., Crawford, M., Debruyn, B., Decaprio, D., Eiglmeier, K., Eisenstadt, E., El-Dorry, H., Gelbart, W. M., Gomes, S. L., Hammond, M., Hannick, L. I., Hogan, J. R., Holmes, M. H., Jaffe, D., Johnston, J. S., Kennedy, R. C., Koo, H., Kravitz, S., Kriventseva, E. V., Kulp, D., Labutti, K., Lee, E., Li, S., Lovin, D. D., Mao, C., Mauceli, E., Menck, C. F. M., Miller, J. R., Montgomery, P., Mori, A., Nascimento, A. L., Naveira, H. F., Nusbaum, C., O'leary, S., Orvis, J., Pertea, M., Quesneville,

H., Reidenbach, K. R., Rogers, Y.-H., Roth, C. W., Schneider, J. R., Schatz, M., Shumway, M., Stanke, M., Stinson, E. O., Tubio, J. M. C., Vanzee, J. P., Verjovski-Almeida, S., Werner, D., White, O., Wyder, S., Zeng, Q., Zhao, Q., Zhao, Y., Hill, C. A., Raikhel, A. S., Soares, M. B., Knudson, D. L., Lee, N. H., Galagan, J., Salzberg, S. L., Paulsen, I. T., Dimopoulos, G., Collins, F. H., Birren, B., Fraser-Liggett, C. M., and Severson, D. W. *Genome sequence of Aedes aegypti, a major arbovirus vector.* **Science (New York, N.Y.)**, 316(5832):1718–23, June 2007. doi: 10.1126/science.1138878.

Paupy, C., Delatte, H., Bagny, L., Corbel, V., and Fontenille, D. *Aedes albopictus, an arbovirus vector: from the darkness to the light.* **Microbes and infection / Institut Pasteur**, 11(14-15):1177–85, December 2009. doi: 10.1016/j.micinf.2009.05.005.

Peischl, S. and Excoffier, L. *Expansion load: recessive mutations and the role of standing genetic variation.* **Molecular ecology**, March 2015. doi: 10.1111/mec.13154.

Poelchau, M. F., Reynolds, J. a., Denlinger, D. L., Elsik, C. G., and Armbruster, P. a. *Transcriptome sequencing as a platform to elucidate molecular components of the diapause response in the Asian tiger mosquito, Aedes albopictus.* **Physiological entomology**, 38(2):173–181, April 2013a. doi: 10.1111/phen.12016.

Poelchau, M. F., Reynolds, J. a., Elsik, C. G., Denlinger, D. L., and Armbruster, P. a. *Deep sequencing reveals complex mechanisms of diapause preparation in the invasive mosquito, Aedes albopictus.* **Proceedings. Biological sciences / The Royal Society**, 280(1759):20130143, January 2013b. doi: 10.1098/rspb.2013.0143.

Poelchau, M. F., Reynolds, J. a., Elsik, C. G., Denlinger, D. L., and Armbruster, P. a. *RNA-Seq reveals early distinctions and late convergence of gene expression between diapause and quiescence in the Asian tiger mosquito, Aedes albopictus.* **The Journal of experimental biology**, 216(Pt 21):4082–4090, November 2013c. doi: 10.1242/jeb.089508.

Porretta, D., Mastrantonio, V., Bellini, R., Somboon, P., and Urbanelli, S. *Glacial history of a modern invader: phylogeography and species distribution modelling of the Asian tiger mosquito Aedes albopictus.* **PloS one**, 7(9): e44515, January 2012. doi: 10.1371/journal.pone.0044515.

Rao, P. N. and Rai, K. S. *Inter and intraspecific variation in nuclear DNA content in Aedes mosquitoes.* **Heredity**, 59 (2):253–258, October 1987. doi: 10.1038/hdy.1987.120.

Santolamazza, F., Mancini, E., Simard, F., Qi, Y., Tu, Z., and della Torre, A. *Insertion polymorphisms of SINE200 retrotransposons within speciation islands of Anopheles gambiae molecular forms.* **Malaria journal**, 7(1):163, January 2008. doi: 10.1186/1475-2875-7-163.

Stapley, J., Santure, A. W., and Dennis, S. R. *Transposable elements as agents of rapid adaptation may explain the genetic paradox of invasive species.* **Molecular ecology**, 24(9):2241–52, May 2015. doi: 10.1111/mec.13089.

Takumi, K., Scholte, E.-J., Braks, M., Reusken, C., Avenell, D., and Medlock, J. M. *Introduction, scenarios for establishment and seasonal activity of Aedes albopictus in The Netherlands.* **Vector borne and zoonotic diseases (Larchmont, N.Y.)**, 9(2):191–6, April 2009. doi: 10.1089/vbz.2008.0038.

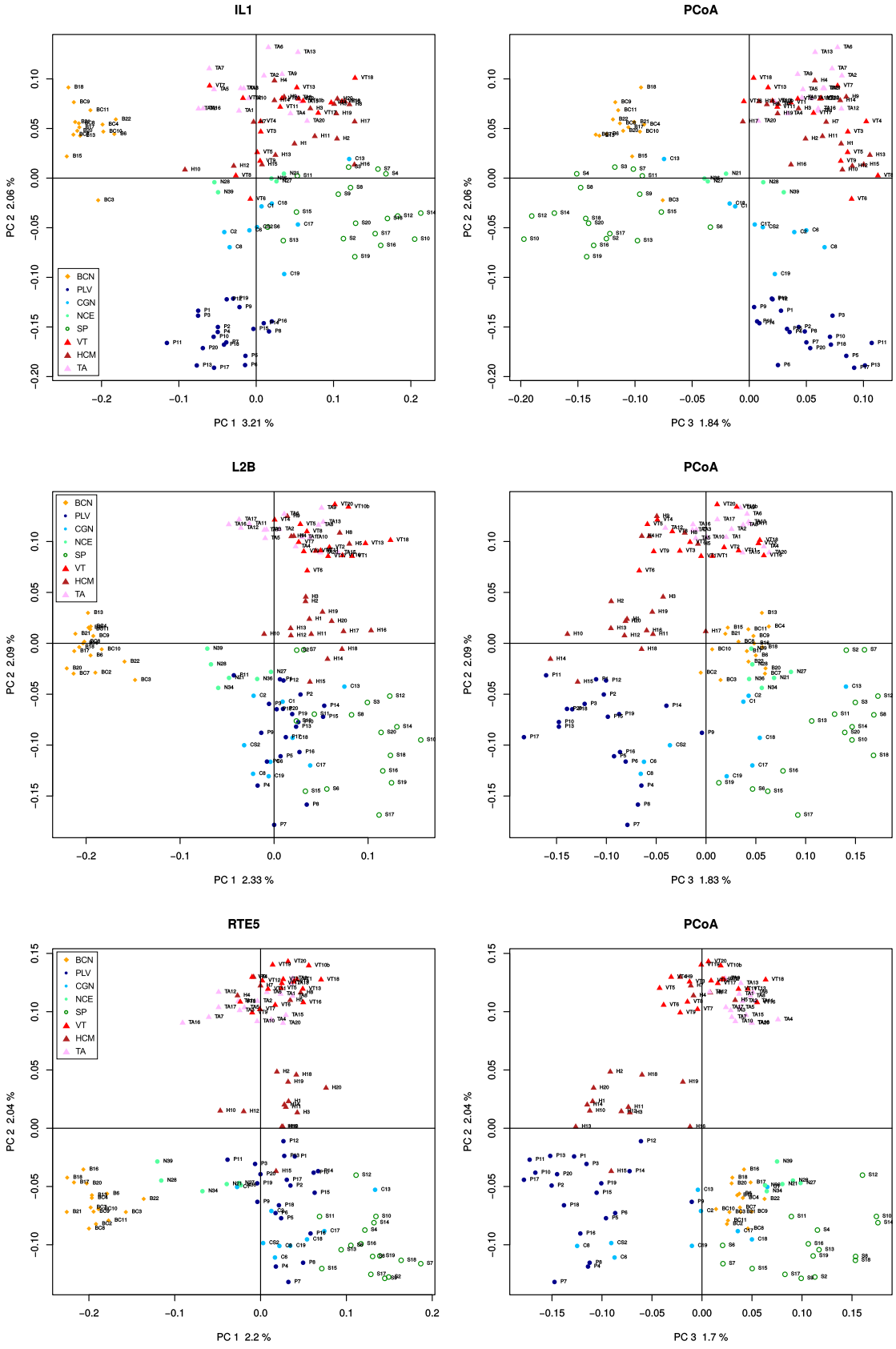
Urbanelli, S., Bellini, R., Carrieri, M., Sallicandro, P., and Celli, G. *Population structure of Aedes albopictus (Skuse): the mosquito which is colonizing Mediterranean countries.* **Heredity**, 84 ( Pt 3)(November 1999):331–337, March 2000.

Zhong, D., Lo, E., Hu, R., Metzger, M. E., Cummings, R., Bonizzoni, M., Fujioka, K. K., Sorvillo, T. E., Kluh, S., Healy, S. P., Fredregill, C., Kramer, V. L., Chen, X., and Yan, G. *Genetic analysis of invasive Aedes albopictus populations in Los Angeles County, California and its potential public health impact.* **PloS one**, 8(7):e68586, January 2013. doi: 10.1371/journal.pone.0068586.

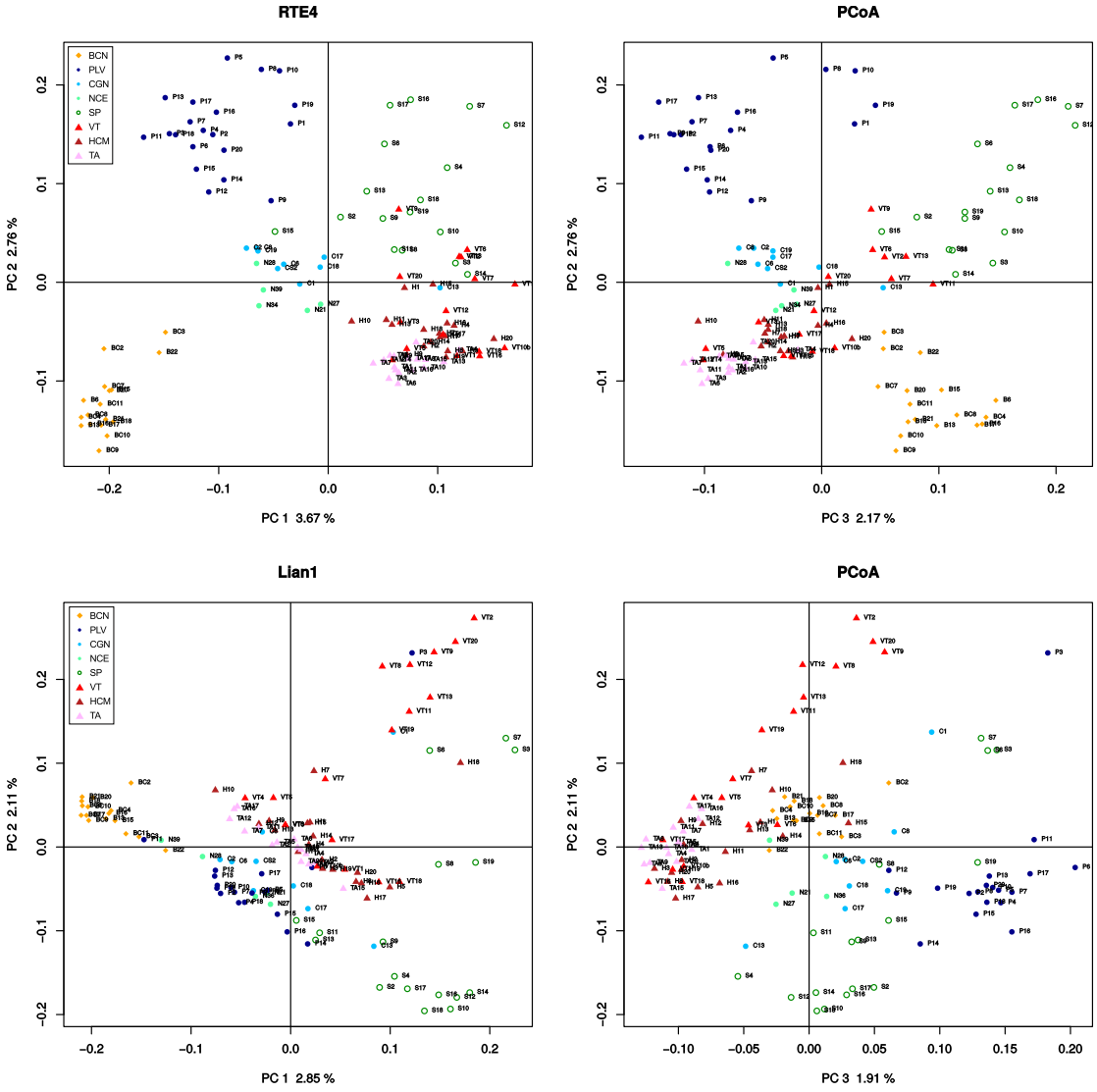
# Supplementary Material

## PCoAs

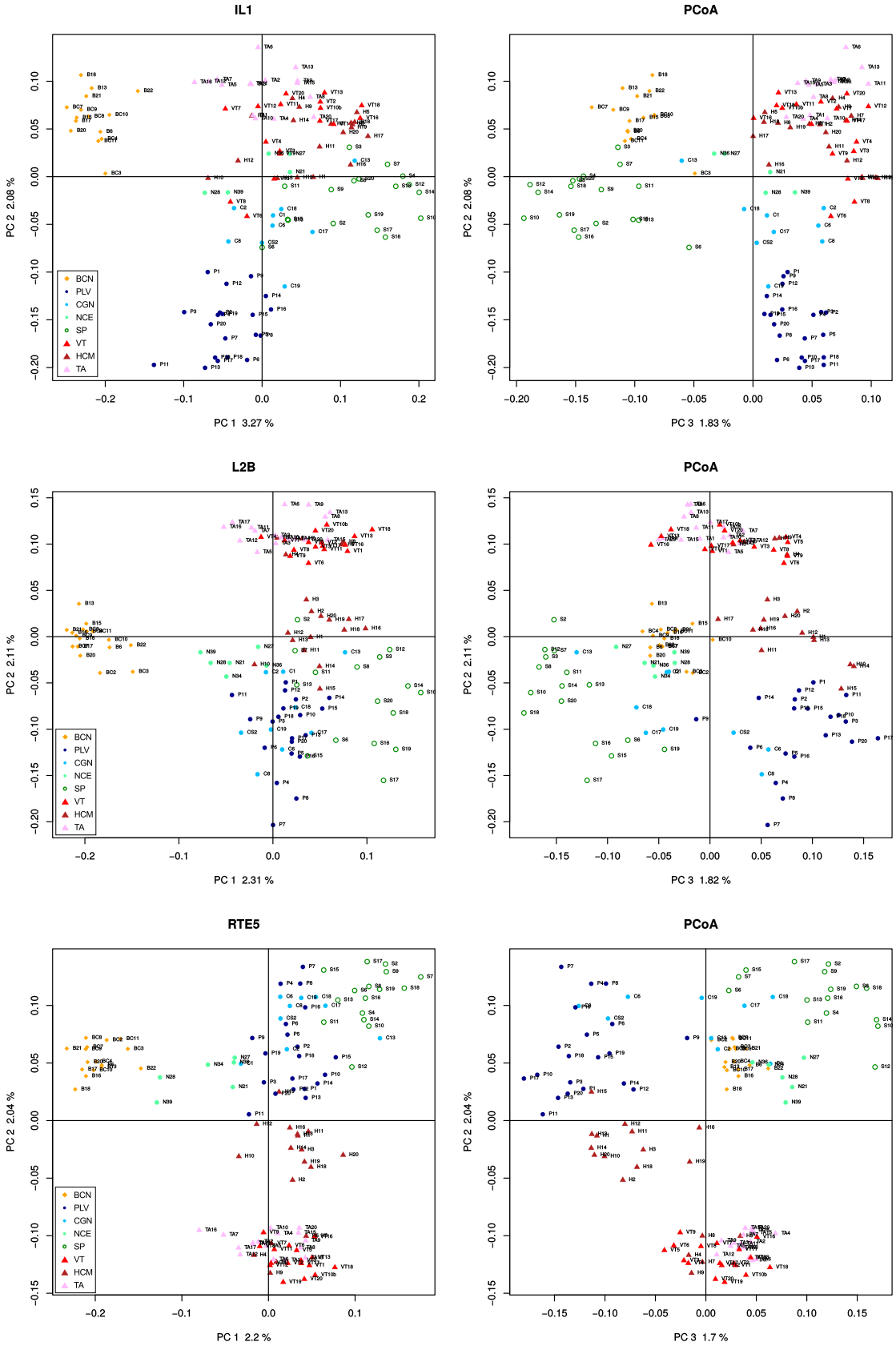
Next page are the PCoAs analysis for M2 and M3 replicates of read sampling.



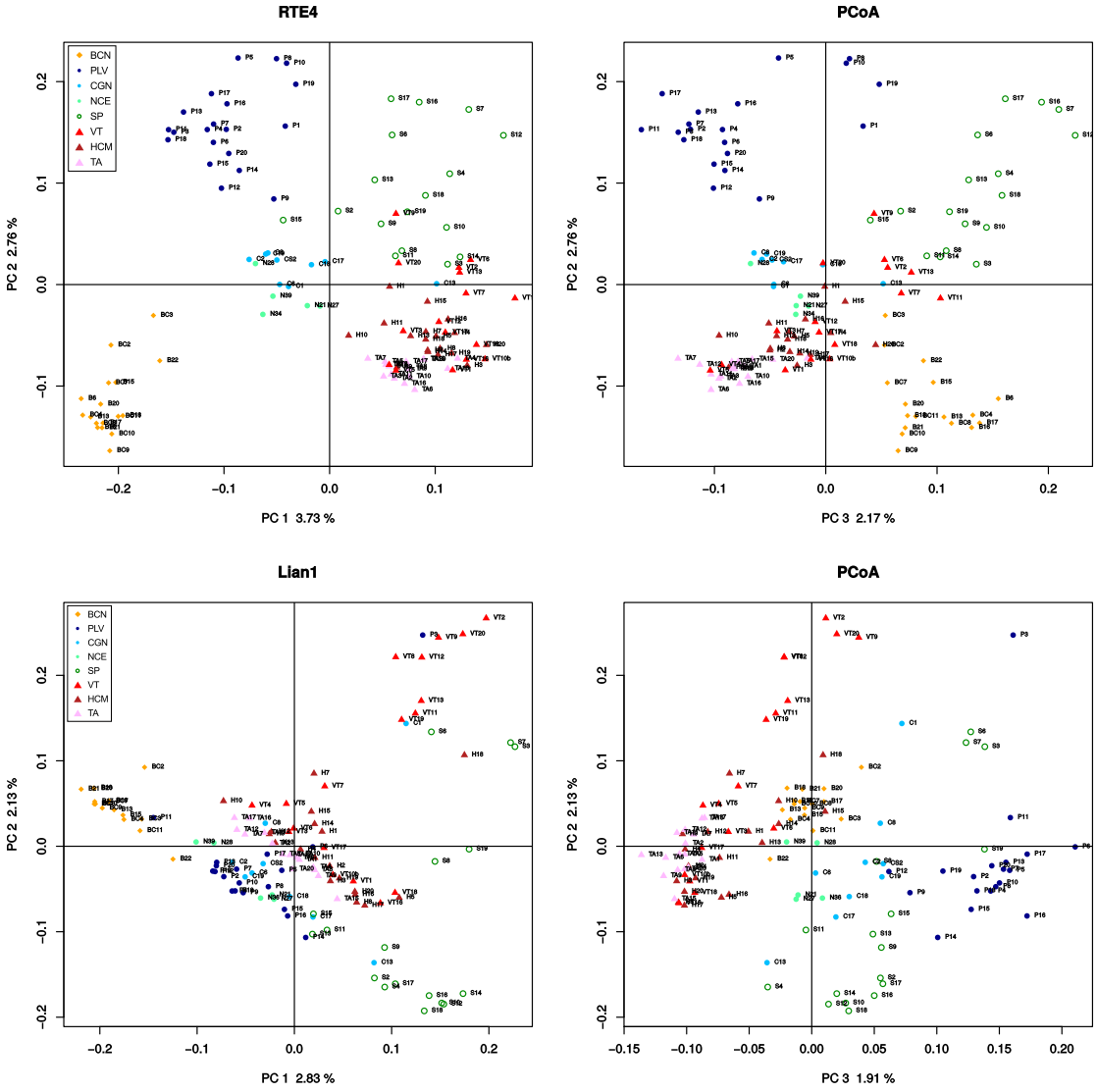

Suppl. Figure S 1




Suppl. Figure S 2




Suppl. Figure S 3

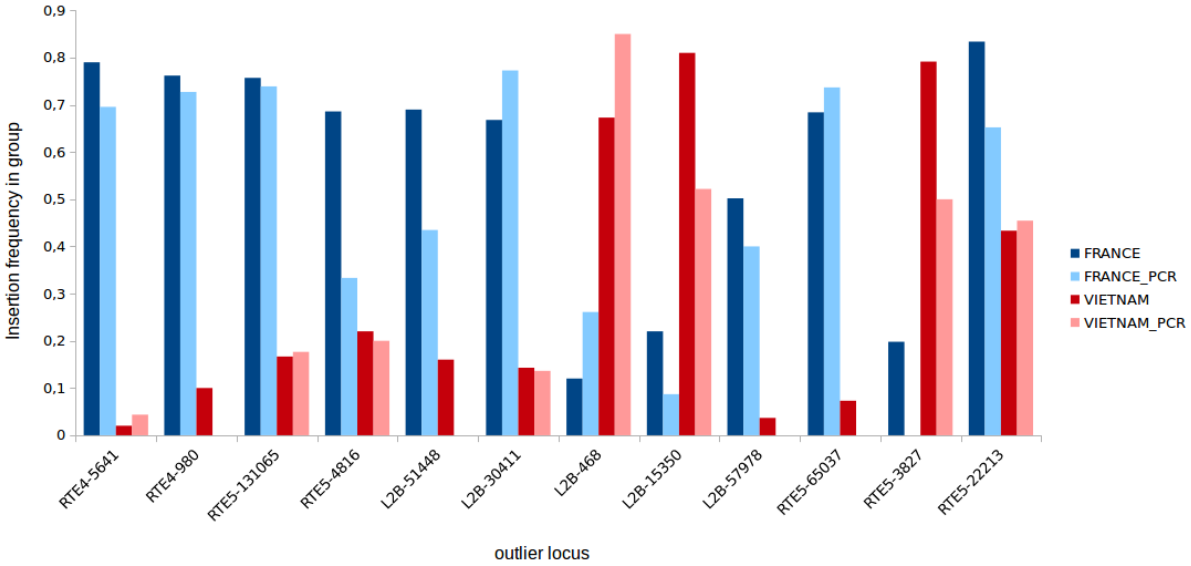



Suppl. Figure S 4

## Pairwise $F_{ST}$

Estimate of pairwise $F_{ST}$ for the three replicates of read sampling are available at the following address: `ftp://pbil.univ-lyon1.fr/pub/divers/goubert/GS_supp/Fst_pairwise`. For each TE familly the first table is the $F_{ST}$ estimate and the second the pairwise $P$-value of $F_{ST}$ estimate. Populations names are: 1=BCN; 2=CGN; 3=NCE; 4=PLB; 5=SP; 6=HCM; 7=TA; 8=VT

## Outlier validation by PCR

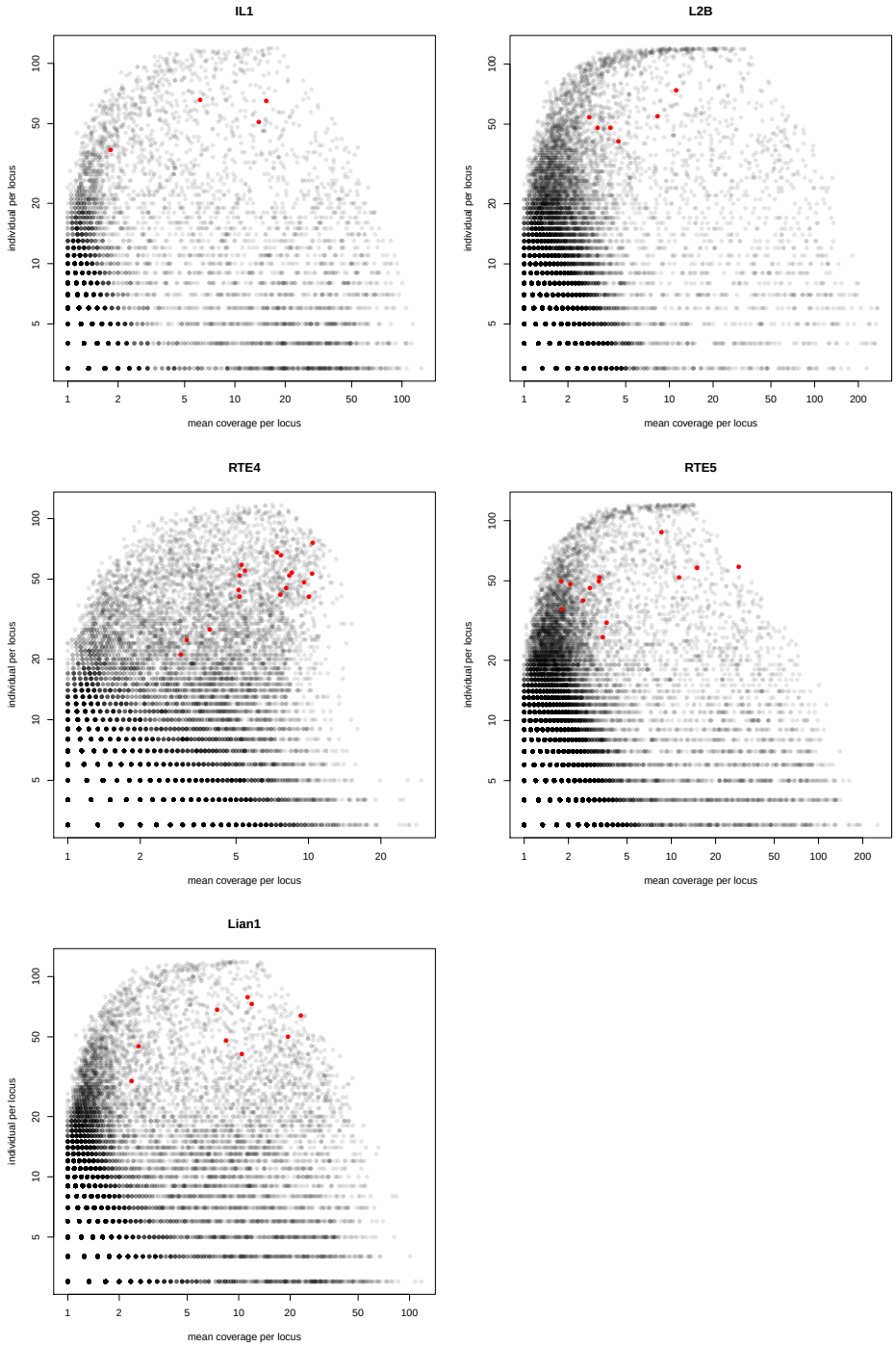



Suppl. Figure S 5 – Insertion frequencies in Europe and Vietnam of 12 outlier loci for which specific PCR was performed. Darker bars are the results from bioinformatic analysis ans lighter are the frequencies obtained over 47 individuals used for PCR validations.

Suppl. Table S 1 – Sequence primer (R) of the 12 outliers used for PCR validation. The corresponding F primer are those given in Table 2 according to the outlier name : L2B = L2B (1/2) F; RTE5 = RTE5 (1/2) F; RTE4 = RTE4 F

| Outlier | Primer sequence 5'-3' |
|---|---|
| RTE5_131065 | GTGGGGCATGGTGCTAAG |
| RTE5_4816 | CAGCCGTTAACACTTTGAGC |
| L2B_51448 | TGCAAAAAGGTAATGGGATTTC |
| L2B_30411 | TCAGGCACCAAGCTACATCA |
| L2B_468 | TAAGGGACGTTTTCCTCTGG |
| L2B_15350 | GCCAAAAACCATGCAGAAAT |
| L2B_57978 | TTGGAGGCACATTCTAGTAGTCA |
| RTE5_65037 | TGGGGCATGGTGCTAAGTAC |
| RTE5_3827 | ACCCTTTGGAAACCCTTGAA |
| RTE5_22213 | CAGATCGGGTACTTCAACTGC |
| RTE4_5641 | TCTCACGACTATTACAGGCATTTT |
| RTE4_980 | CACTGAATTTCTTTTCCTTGGAA |

## Relationship between mean locus coverage and insertion frequency




Suppl. Figure S 6 – Relationship between mean locus coverage (mean number of read per individual at one locus when present) and insertion frequency (number of individuals that share the locus) for the M1 replicate (others have simimlar results). Red dots are outlier loci.

## Indexes used for library multiplexing

Suppl. Table S 2 – Individual indexes used for multiplexing

| Index name | Sequence 5'-3' |
|---|---|
| A001 | GACGAT |
| A002 | ATTCTC |
| A003 | CGAAGG |
| A004 | TTGAAG |
| A005 | AAAGTA |
| A006 | TCAGCG |
| A007 | AGGCGC |
| A008 | AATCCG |
| A009 | TGATGC |
| A010 | ATCTAT |
| A011 | TAGATC |
| A012 | GTGAAT |
| A013 | AACGTC |
| A014 | TGTCGT |
| A015 | CACTAA |
| A016 | TTGACT |
| A017 | ACATTA |
| A018 | GGGTCT |
| A019 | CCAAGC |
| A020 | GCACTG |
| A021 | CGTTAA |
| A022 | GTGTAG |
| A023 | CACCTC |
| A024 | TAAATG |
| A025 | CTTGAC |
| A026 | TTAAGG |
| A027 | AAGTAA |
| A028 | GCTTCG |
| A029 | AGAATC |
| A030 | GACTTT |
| A031 | CGGAAC |
| A032 | ATTTGT |
| A033 | GCCCAC |
| A034 | AACCAT |
| A035 | GGGGGA |
| A036 | TAGTAC |
| A037 | GTTCTT |
| A038 | ACAGAC |
| A039 | GTTGCG |
| A040 | AAACTC |
| A041 | AGAAGT |
| A042 | TCTTAA |
| A043 | CGCGTG |
| A044 | GCGCCC |
| A045 | AAGGCC |
| A046 | TGCACG |
| A047 | CAGTGA |
| A048 | GCCGGT |
| A049 | CTTAAA |
| A050 | TAAGGG |
| A051 | ATCGTT |
| A052 | GAGCAA |
| A053 | ACTGCG |
| A054 | GGCCTC |
| A055 | ATCGGG |
| A056 | AGTCGG |

| | |
|---|---|
| A057 | TACTCA |
| A058 | ATAACG |
| A059 | GAGGGC |
| A060 | TACAAG |
| A061 | CGTTTC |
| A062 | CCCGTT |
| A063 | GGTCAC |
| A064 | AGTGCT |
| A065 | GACATC |
| A066 | CCCGGC |
| A067 | CTCGGT |
| A068 | TCGAAC |
| A069 | GTGTTT |
| A070 | ACCCCC |
| A071 | ACTTTT |
| A072 | GGCCAA |
| A073 | GACAGT |
| A074 | ATGTCA |
| A075 | GCAGCT |
| A076 | CGTCGC |
| A077 | CGTTGG |
| A078 | ACCAGG |
| A079 | TGTGCC |
| A080 | GGACTT |
| A081 | CCGTCA |
| A082 | CATCCT |
| A083 | GTCGGC |
| A084 | TGTAAA |
| A097 | CTACTA |
| A098 | TGCGCT |
| A099 | ACGATC |
| A100 | CGCTCT |
| A101 | GATAGA |
| A102 | TATCAT |
| A103 | CTAGTC |
| A104 | GGCTTG |
| A105 | CCTCCC |
| A106 | GCACGT |
| A107 | AGGGCA |
| A108 | TCCAGA |
| A109 | GATCTG |
| A110 | CGCGCC |
| A111 | GCCGCG |
| A112 | TAGGAA |
| A113 | TATCGA |
| A114 | TCGAGG |
| A115 | CGATAC |
| A116 | CACCGG |
| A117 | TGGTCA |
| A118 | TCATGT |
| A119 | TCTCTC |
| A120 | GTAGTT |
| A121 | CAGAAA |
| A122 | AGTAAG |
| A123 | TCCTTA |
| A124 | CTTGGG |
| A125 | GAACAC |
| A126 | CGGCAT |
| A127 | TACGGC |
| A128 | GGTTAT |
| A129 | TCTGAT |

| | |
|---|---|
| A130 | CTCATA |
| A131 | GAGTTG |
| A132 | CGCACA |
| A133 | AATTCT |
| A134 | GGAAGA |
| A135 | CTAGGA |
| A136 | TGACCT |
| A137 | CAGTTC |
| A138 | TGCAGT |
| A139 | AATGAA |
| A140 | TGCCAG |
| A141 | CACTGT |
| A142 | GTGCCA |
| A143 | AAATGT |
| A144 | GTGACG |
| A145 | GGTCCG |
| A146 | AACTTA |
| A147 | TGTGAG |
| A148 | CAACGA |
| A149 | AAGGTG |
| A150 | TTCAAC |
| A151 | ACGAAT |
| A152 | GGATTC |

# Discussion Générale

*"I find your lack of faith disturbing"*

– Darth Vader, *Star Wars IV : A new hope* 1977

Parmi les moyens ayant conduit à l'adéquation des espèces invasives avec leurs nouveaux environnements, l'adaptation a souvent été évoquée comme un processus fréquent, mais cette affirmation se doit d'être étayée par l'apport de résultats empiriques (Handley *et al.* 2011 ; Savolainen *et al.* 2013 ; Colautti et Lau 2015).

Le moustique tigre suscite aujourd'hui un grand intérêt du fait de son importance épidémiologique et la rapidité avec laquelle il a colonisé de nombreuses régions de par le monde. Son aire de répartition actuelle inclut des environnements aussi contrastés que peuvent l'être les forêts humides d'Asie du sud-est et les grandes villes du nord des États-Unis, ce qui démontre, comme nous avons pu l'évoquer, une grande versatilité de la niche écologique de cette espèce.

Au cours de cette thèse, notre principal objectif a été de rechercher d'éventuelles signatures laissées par la sélection naturelle sur les génomes, sans *a priori* sur les régions concernées et les traits qui pourraient y être associés. Pour cela, nous nous sommes intéressés à la différenciation génétique entre un groupe de populations tropicales au Vietnam et un groupe de populations invasives en Europe, installées dans un environnement tempéré. La poursuite de cet objectif a nécessité la mise en œuvre de nouvelles méthodes bio-informatiques et de biologie moléculaire, nous permettant d'accéder à des marqueurs génétiques adaptés à la réalisation d'un scan génomique.

Le résultat principal de ces travaux est la mise en évidence de 92 locus *outliers*, qui pourraient être dus a des évènements d'adaptation, principalement observés au sein des populations tempérées. Ces recherches ont par ailleurs été l'occasion, d'enrichir nos connaissances sur la génomique de cette espèce, notamment quant à la taille de son génome, sa composition en ADN répété et la dynamique de ses éléments transposables. *Ae. albopictus* est en effet une espèce qui se distingue par ces traits d'un grand nombre des insectes dont le génome est décrit aujourd'hui.

# Développement de méthodes originales

## `dnaPipeTE`, l'histoire d'un détournement

Afin de développer de nouveaux marqueurs génétiques basés sur le polymorphisme d'insertion des éléments transposables chez *Ae. albopictus*, nous devions dans un premier temps identifier chez cette espèce des familles d'ET très répétées, dont les différentes copies sont suffisamment proches, pour pouvoir amplifier avec la même amorce le plus grand nombre d'insertions possible. Le principal défi consistait à réaliser cette analyse sans génome de référence. Nous disposions cependant des lectures non assemblées, produites par le séquençage à haut débit (sur plateforme Illumina) d'individus d'une souche isolée sur l'île de La Réunion. Parmi les approches envisagées, l'assemblage de l'ADN répété basé sur leur sur-représentation au sein des lectures brutes nous a semblé être une

méthode pertinente, et des outils dédiés, en particulier le pipeline `RepeatExplorer` (RE, Novák *et al.* 2010), nous ont permis de rapidement cibler des familles correspondant à nos attentes.

Un résultat important de ces primo-analyses a été de découvrir que la moitié du génome de cette souche était composé d'ADN répété, et il nous est apparu important pour cette espèce qui dispose, encore à l'heure actuelle, de peu de ressources génomiques (un transcrtiptome annoté (Poelchau *et al.* 2013a) ; et très récemment un génome "brut" extrêmement fragmenté (Dritsou *et al.* 2015)) d'apporter une description exhaustive de cette fraction importante du génome d'*Ae. albopictus*. Comme cela est évoqué dans le chapitre 2, le pipeline `RepeatExplorer` a rapidement présenté plusieurs limitations, qu'il nous a semblé possible de dépasser tout en maintenant l'idée originale de mener un assemblage spécifique du répétome en utilisant seulement une fraction inférieure à une fois (1X) la taille du génome des lectures de séquençage.

Parmi les améliorations possibles, la méthode d'assemblage à proprement parler pouvait être revue. RE permet suite à une comparaison deux à deux de lectures de l'échantillon de visualiser sous forme de graphique leurs connexions les unes avec les autres au sein d'un "cluster" qui correspond en théorie à une famille de répétitions. La projection graphique d'un cluster permet par exemple d'identifier différents "chemins" si des variants de structure (insertions, délétions) existent entre les différentes copies d'une famille. Cette information est très utile, et constitue une innovation certaine par rapport aux méthodes préexistantes (voir Modolo et Lerat 2014). Cependant, l'assemblage des différents variants structuraux demeure une étape manuelle : les lectures de chaque cluster sont assemblées par RE avec l'outil CAP3, qui n'est pas conçu pour prendre en charge les reads courts produits par les séquenceurs de type NGS ; en effet, il ne prend pas en compte l'information sur les possibles structures alternatives : il assemble donc les différents chemins de manière indépendante et produit plusieurs contigs qu'il sera nécessaire d'assembler manuellement. Il devient alors plus difficile d'identifier sur la séquence assemblée des motifs ou des régions conservées permettant d'annoter la famille de répétition.

En détournant l'usage de l'assembleur `Trinity` (Grabherr *et al.* 2011), `dnaPipeTE` remplace le clustering des lectures par Blastn et l'assemblage des clusters par CAP3 par une méthode dédiée à l'assemblage des séquences courtes mais aussi à la gestion des séquences alternatives. En effet, étant au départ développé pour l'assemblage sans référence des transcrits (RNA-seq), `Trinity` est conçu pour gérer les fortes variations de couverture, qui fluctue en RNA-seq avec le niveau d'expression d'un gène (qui devient dans notre cas le nombre de copies d'une séquence répétée), et les transcrits alternatifs (qui sont pour nous équivalents aux variants structuraux). Les résultats obtenus avec `dnaPipeTE` sont très concluants et améliorent considérablement les métriques d'assemblage (voir table 2 de l'article 2). Il est par exemple possible de retrouver la séquence complète de deux variants de la même famille parmi les ET les plus abondants chez *Ae. albopictus*, en

utilisant seulement 0,3X de son génome. L'utilisation de `Trinity` permet par ailleurs à la méthode de `dnaPipeTE` de gagner plusieurs heures voire jours par rapport au pipeline RE.

D'autres améliorations concernent l'annotation automatique des répétitions, réalisée en utilisant le logiciel `Repeat Masker` (RM) directement sur les contigs produits par `dnaPipeTE`, et non sur les reads, comme c'est le cas avec RE. Cette procédure a permis d'identifier automatiquement, au niveau de leur clade, jusqu'à 85,3% des répétitions présentes dans le génome du moustique tigre, contre 25,5% avec RE, où cette étape doit être faite manuellement.

Enfin, `dnaPipeTE` est la première méthode permettant d'estimer l'âge relatif des différentes familles d'ET présentes au sein d'un génome à partir de lectures non assemblées. Jusqu'à présent une telle analyse était possible en réalisant un alignement nucléotidique des différentes copies d'une même famille, préalablement identifiées à l'aide de RM au sein d'un génome assemblé. La divergence moyenne entre ces copies peut être interprétée comme une estimation du temps écoulé depuis leur transposition, en faisant l'hypothèse d'une accumulation régulière de mutations aléatoires au cours du temps (horloge moléculaire). De manière analogue, lorsque des reads appartenant à différentes copies d'un ET sont mappées par `dnaPipeTE` sur leur contig de référence, la distribution des valeurs de divergence nucléotidique entre reads et contigs est relative au nombre de mutations accumulées par les différentes copies, et sous la même hypothèse que précédemment, au temps. La comparaison de nos résultats chez différentes espèces avec les profils obtenus en utilisant RM sur des génomes assemblés est très convaincante (voir Figure S3 du matériel supplémentaire de l'article 2). Pour être complet, il convient de noter que la méthode RE (Novák *et al.* 2010) permet d'avoir une idée de la divergence entre les différentes copies de chaque famille d'ET. En effet, la représentation graphique de la distance nucléotidique entre les lectures appartenant à un même cluster, permet de distinguer des familles dont les arêtes qui relient les reads entre eux sont plus longues (leur taille est proportionnelle à la distance nucléotidique entre deux reads qui se superposent) mais aussi plus souvent associées à des reads uniques ce qui s'observe lorsque les copies dont proviennent les reads sont très divergentes. Cependant, RE n'inclut aucun outil permettant de comparer ces métriques (ni même de les récupérer automatiquement) entre clusters et donc d'effectuer directement une analyse globale de la dynamique des ET dans l'échantillon analysé.

La principale limite de la méthode est la même pour toutes celles basées sur l'assemblage des répétitions à partir d'un séquençage à faible couverture : les répétitions les plus fréquentes, qui auront donc plus de lectures, seront les mieux assemblées. Une partie du répétome peut ainsi être sous-estimée s'il est composé de familles peu répétées. D'autre part, l'analyse est restreinte aux familles dont les copies sont suffisamment proches pour être assemblées entre elles. Certains génomes, qui abritent comme chez l'Homme de nombreuses familles depuis longtemps inactives, se prêtent moins bien à ce genre d'analyse :

la proportion du répétome est alors sous estimée (voir le paragraphe 1.3 du matériel supplémentaire de l'article 2 et la figure S4). La même limite a par ailleurs pu être constatée avec l'utilisation du pipeline RE sur les mêmes données. Enfin, ce manque de sensibilité rend difficile l'observation des familles les plus anciennes lors de l'analyse de l'âge relatif des différentes familles, en témoigne l'absence des ET probablement les plus anciens des graphiques obtenus avec `dnaPipeTE` lorsqu'ils sont comparés à ceux des "landscapes" obtenus après l'analyse d'un génome complet avec RM. Ainsi, il convient d'être prudent lors de l'interprétation des résultats, car la présence ou non de familles "anciennes" ne peut être évaluée, sauf de manière comparative entre différents échantillons.

`dnaPipeTE` permet tout de même d'obtenir rapidement, à partir de petits échantillons, une estimation de la quantité, de la diversité et de la dynamique des ET. Ce pipeline permet d'envisager des analyses comparatives (telle celle réalisée entre *Ae. albopictus* et *Ae. aegypti* dans cette thèse) basées sur du séquençage à faible couverture, ce qui réduit les coûts au profit d'une comparaison d'un plus grand nombre d'échantillons. La méthode a vocation à devenir un outil à disposition de la communauté intéressée dans l'analyse comparative des ET et permet d'obtenir des informations (estimation du contenu, contigs de référence) utiles lors de l'assemblage de nouveaux génomes. La méthode fait aujourd'hui partie intégrante de projets indépendants développés au sein de notre équipe au laboratoire, et sa publication a suscité l'intérêt de différentes équipes à l'étranger.

Au titre des améliorations prévues, l'intégration des différents scripts et programmes du pipeline dans un package unique ne nécessitant aucune installation particulière devra permettre de rendre plus facile son utilisation. En effet, de tels pipelines nécessitent l'installation manuelle des différentes dépendances (programmes, librairies,...) ; nous avons essayé de rendre cette tache la plus simple possible mais il est encore nécessaire de configurer manuellement certains outils. La solution envisagée est l'intégration du pipeline dans un "conteneur" incluant une machine virtuelle. La pré-installation des programmes et librairies étant alors réalisée en amont, l'utilisateur n'aura qu'à télécharger le conteneur (cependant plus lourd que les simples scripts du pipeline) et pourra réaliser directement les analyses.

La conception du pipeline, sous la forme d'un script principal (open source) et l'utilisation de différents programmes le rend très facile à modifier en fonction des différents intérêts. Ainsi, seront très rapidement intégrées des améliorations suggérées par les premiers utilisateurs, notamment l'intégration des banques de données ET utilisées par RM, une option facilitant l'utilisation d'une banque personnelle d'ET ou le calcul automatique du nombre de lectures nécessaires en fonction de la longueur des reads, la couverture souhaitée et la taille supposée du génome (calcul aujourd'hui réalisé par l'utilisateur). Sur le plus long terme, il serait intéressant de tester d'autres assembleurs, et notamment ceux spécialement développés pour les éléments transposables comme TEDna (Zytnicki *et al.* 2014) et REPark (Koch *et al.* 2014). A la différence de la méthode actuelle, ces

assembleurs se basent sur la sur-représentation des $k$-mer (mots de $k$ lettres composant les lectures) dans l'échantillon total qui sont alors spécifiquement assemblés. Si leur utilisation nécessitera donc un séquençage plus profond qu'actuellement, elle peut être vue comme une version alternative, complétant le champ des applications du pipeline, notamment concernant l'assemblage et l'annotation des répétitions lors des projets d'assemblage de génomes complets.

## Le Transposon Display à haut débit

Une fois les 5 familles d'ET utilisées pour le scan génomique identifiées et validées expérimentalement, notre "cahier des charges" était d'arriver à séquencer pour plus d'une centaine d'individus, plusieurs milliers d'insertions, puis de retrouver par analyse bioinformatique chacun de ces locus parmi l'ensemble des paires de lectures de chacun des moustiques, sans passer par leur alignement sur un génome de référence.

De telles études du polymorphisme d'insertion des ET ont été principalement menées chez des espèces modèles (Homme, riz, fraise, levure), qui au contraire disposent certainement des meilleures ressources génomiques possibles (Iskow *et al.* 2010 ; Witherspoon *et al.* 2010 ; Sabot *et al.* 2011 ; Bridier-Nahmias *et al.* 2015 ; Monden et Tahara 2015). L'approche avec génome de référence a en effet plusieurs avantages : elle permet tout d'abord de s'assurer qu'un locus d'insertion identifié correspond à une partie réelle du génome étudié, et peut alors permettre de connaître sa localisation précise, ce qui est alors utile à l'interprétation des résultats du scan génomique. D'autre part, l'ancrage des insertions à des positions uniques d'un génome permet de s'assurer que le polymorphisme observé entre individus concerne bien la même insertion.

Récemment, une méthode très similaire à la notre a été développée afin d'étudier sans génome de référence, le polymorphisme de 2024 insertions de deux familles d'ET, parmi 38 cultivars de la patate douce (Monden *et al.* 2014). Dans notre étude, nous avons montré qu'il était possible d'identifier plus de 128 000 locus d'insertions en procédant au clustering des lectures correspondant aux régions flanquantes de chacune des familles d'éléments. Cette procédure sans génome de référence est en fait assez analogue à celle utilisée pour le RAD-seq, qui est comme nous l'avons vu aujourd'hui fréquemment utilisé chez des espèces non-modèles : dans ce cas, les RAD-tags séquencés sont eux aussi comparés uns à uns afin de retrouver les différents locus. Nous avons dû développer une procédure informatique sur mesure afin de traiter nos données, permettant d'automatiser les différentes étapes, du filtrage des séquences brutes aux analyses de génétique des populations. Cette phase de développement a notamment permis d'explorer différents jeux de paramètres afin d'évaluer la robustesse des résultats par rapport à ces fluctuations.

Notre principale préoccupation a été de gérer la variation de couverture de séquençage entre les individus, la plus grande différence étant d'environ 10X entre deux individus

(≈200 000 - 2 000 000 de reads). Ces variations pourraient être liées à la normalisation des banques de séquençage, mais nous n'avons cependant pas trouvé de corrélation entre la quantité d'ADN purifié (avant normalisation) et le nombre de lectures obtenues au final. La présence d'une telle hétérogénéité est donc à prévoir en cas de renouvellement de l'expérience, et une attention particulière devra être apportée aux concentrations initiales. Il pourrait être par exemple informatif d'inclure des réplicats des mêmes produits de TD à différentes concentrations initiales pour un individu donné, afin de voir quelles sont les effets de ces variations sur le patron de présence/absence qui lui est attribué.

Nous avons malgré tout réussi à nous affranchir de cet écueil en réalisant un ré-échantillonnage des lectures au sein du jeux de données initial. Certaines populations européennes étaient en effet initialement mieux couvertes en moyenne que les populations vietnamiennes ; sans correction, nous aurions pu introduire une différenciation artificielle entre les continents, du fait d'un plus grand nombre de faux négatifs (insertion réelle mais non détectée par manque de couverture) au sein des populations asiatiques. Le ré-échantillonnage a alors permis de répartir les éventuels faux négatifs de manière équitable entre l'ensemble des populations. Nous avons observé une homogénéité des résultats entre les différents réplicats de cette procédure, et la comparaison des résultats finaux avec les données non échantillonnées a montré que la structure des populations était en fait très robuste aux variations de couverture. Cette constance est certainement liée au fait qu'un très grand nombre de locus ($\geq$ 10 000 par famille d'ET) ont été analysés. Il n'en demeure que maintenir les variations initiales de couverture aurait certainement contribué à introduire des *outliers* faux positifs pour la sélection (si une faible fraction des locus avaient présenté une différenciation artificielle, ils auraient pu en effet être détectés par le scan génomique mais n'avoir que peu d'influence sur la structure globale des populations).

En appliquant notre procédure de ré-échantillonnage, l'incertitude quant à la quantité de locus d'insertion découverte est répartie de manière homogène sur l'ensemble des individus. Le risque de cette procédure, aurait pu être de baisser artificiellement les fréquences d'insertion globales. Cependant nous avons pu constater que la fréquence d'insertion au sein des populations n'était pas corrélée à la couverture des individus. Autrement dit, les insertions les mieux couvertes ne sont pas forcément celles présentes à fortes fréquences. (C.f. Matériels supplémentaires de l'article 3).

Enfin, l'amplification de chacun des outliers a été systématiquement testée par PCR. Parmi les 92 locus candidats, 12 ont pu être amplifiés directement avec un protocole standardisé. Plutôt que d'insister sur chacun des outliers, nous avons préféré tester pour ces 12 marqueurs le patron de polymorphisme observé par PCR à celui obtenu lors du génotypage par séquençage, et ces expériences ont à chaque fois confirmé le patron observé (voir matériel supplémentaire de l'article 3). Cette expérience de TD à haut débit est par ailleurs la première à introduire la procédure de multiplexage développée par nos partenaires du GénoToul (plateforme publique de génomique basée à Toulouse). Afin

d'incorporer un identifiant unique à chaque individu, les amorces ET utilisées lors de l'amplification des insertions comportaient une séquence commune, présente en 5'.

Cette séquence est ensuite utilisée pour hybrider par PCR une construction comprenant l'adaptateur Illumina (servant à la fixation de la molécule sur la Flowcell du séquenceur) et l'index propre à chaque individu. Cette méthode permet de réduire les coûts de multiplexage à 5€ [1] par individu, et a permis dans notre cas de séquencer sur la même ligne 140 échantillons différents.

L'ADN répété, et en particulier les ET, peut être un souci important lors du développement des méthodes comme le RAD-seq (homoplasie des RAD tags présents dans les séquences répétées, Catchen *et al.* 2011 ; Hohenlohe *et al.* 2013). Le TD à haut débit peut donc être envisagé comme une réelle solution alternative afin de générer facilement plusieurs milliers de marqueurs polymorphes, sans connaissances avancées sur le génome d'une espèce. La forte analogie du TD avec les AFLP permet en outre de profiter de la large collection de logiciels et méthodes développées précédemment afin d'utiliser ces marqueurs dominants.
Ces développements méthodologiques, initiés pour la mise en place du scan génomique, ont été aussi à la base d'analyses originales, contribuant aux connaissances sur la génomique d'*Ae. albopictus*.

## Un génome en pleine invasion ?

Le génome du moustique tigre est remarquable par sa taille et la proportion d'ADN répété qu'il contient. Notre analyse d'une souche isolée à la Réunion, pour laquelle nous avons estimé la taille du génome à 1,16 Gpb, a révélé qu'un tiers au moins était composé d'ET, et que la fraction d'ADN répété atteignait 50% du génome. Une telle taille est cohérente avec les précédentes estimations faites chez cette espèce (Black et Rai 1988 ; Kumar et Rai 1990), ainsi que celles disponibles pour plusieurs autres moustiques du genre *Aedes*, chez qui la taille estimée du génome est souvent proche voir supérieure à 1 Gpb (Gregory, 2005. Animal Genome Size Database. http ://www.genomesize.com.). Ce qui apparaît remarquable, notamment suite à l'analyse comparative avec *Ae. aegypti*, est l'état de très forte conservation des différentes copies d'ET de Classe I qui représentent chez l'une et l'autre des souches étudiées, près de la moitié des ET détectés.

Plusieurs indices nous laissent à penser que ces éléments pourraient être, ou avoir été récemment, très actifs au sein de ces génomes. En plus du bon état de conservation des copies, nous avons montré chez *Ae. albopictus* qu'un très grand nombre des éléments identifiés étaient présents dans le transcriptome de référence, qui inclut les séquences provenant d'individus entiers, aux différents stades œuf, larve et femelle adulte (Armbruster et Poel-

1. Si cette thèse est lue dans un futur lointain, sachez qu'avec 5€ nous pouvons acheter 5 baguettes de pain, un sandwich + frites ou encore 14 cafés au distributeur du laboratoire

chau : http ://www.albopictusexpression.org/). D'autre part, si les proportions globales des différentes classes de répétitions apparaissent bien conservées entre les deux espèces, certaines familles d'ET communes aux deux espèces se retrouvent dans des proportions très différentes, signe que des amplifications majeures ont pu se produire récemment dans l'histoire de ces espèces. Enfin, notre étude de TD basée sur le polymorphisme d'insertion de 5 familles de LINE a révélé un très fort polymorphisme d'insertion parmi plus d'une centaines d'individus. La plupart de ces insertions – plus de 128 000 identifiées – se trouvent en effet à très faible fréquence au sein des populations étudiées (1 insertion est en moyenne partagée par une dizaine d'individus, mais la médiane de cette distribution est elle proche de 5 individus). Ces éléments, qui sont les plus abondants au sein de la souche étudiée dans le chapitre 3, pourraient donc avoir eu une activité de transposition très récente, voire être toujours actifs au sein de l'espèce.

Ces résultats sont à mettre en parallèle avec les travaux menés à la fin des années 1980 sur la taille du génome de plusieurs populations d'*Ae. albopictus*. Les variations du simple au double observées à l'époque étaient suspectées d'être liées à des variations du contenu en éléments répétés (McLain *et al.* 1987 ; Black et Rai 1988). Si nos résultats ne permettent pas de valider ces travaux, le faisceau d'indices accumulés aujourd'hui rend plausible ces hypothèses. Il serait particulièrement intéressant de répéter ces expériences avec des méthodes modernes, comme la cytométrie de flux, afin de savoir si ces variations intra-spécifiques de taille de génomes peuvent être confirmées. D'autre part, il est envisageable de concevoir une étude comparative du contenu de ces souches en ET, à partir du séquençage à faible couverture de leurs génomes. L'analyse combinée des tailles des génomes et du contenu en ET avec `dnaPipeTE` pourrait permettre de mieux comprendre ce patron de variation, et éventuellement d'identifier les familles d'éléments responsables de ces variations importantes de tailles de génomes.

Il serait par ailleurs important de voir s'il existe une relation entre la taille du génome, le contenu en ET et la répartition géographique du moustique tigre. En particulier, il serait intéressant de déterminer si certains environnements sont associés à certaines tailles de génome, et à quel point les ET sont impliqués dans ces processus. Bien que les méthodes utilisées à l'époque pour mesurer la taille des génomes invitent aujourd'hui à la prudence, Kumar et Rai (1990) montraient que des souches invasives présentes aux USA et au Brésil possédaient une taille de génome significativement supérieure à celle des populations observée dans l'aire d'origine. La même étude contredisait cependant l'hypothèse de Rao et Rai (1987) qui supposaient que l'insularité de certaines populations (Indonésie, Hong Kong, Japon, Hawaï, Madagascar et Maurice) pouvait être corrélée à une augmentation de la taille du génome, par rapport aux échantillons prélevées dans l'aire d'origine continentale (Inde). S'appuyant sur les résultats de Ferrari et Rai (1989) montrant une corrélation positive entre la taille du génome et le temps de développement chez *Ae. albopictus*, Rao et Rai (1987) proposaient que ce phénotype pourrait avoir un

rôle (cependant inconnu) dans le succès écologique des populations invasives.

Récemment, une étude menée chez *D. melanogaster* a rapporté une variation de 6% du contenu en ET entre des populations présentes dans deux micro-habitats très différents du Mount Carmel, en Israël (Kim *et al.* 2014). Ces variations sont associées à des insertions micro-habitats spécifiques, perturbant des gènes déjà associés à des divergences phénotypiques entre ces populations. Chez *Ae. albopictus*, la question est donc de savoir, si les variations de taille de génomes dues aux ET sont réelles, et le cas échéant, quels processus peuvent être responsables de la mobilisation des ET (stress climatique ou lié au changement d'environnement), et quelles forces évolutives (sélection, dérive) pourraient être à l'origine du polymorphisme d'insertion, et par extension, de la taille des génomes.

## Premier jalon vers la génomique de l'adaptation d'*Ae. albopictus*

Objectif principal de ces travaux de thèse, le scan génomique réalisé chez *Ae. albopictus* est la première étude cherchant à mettre en évidence des bases génétiques liées à l'adaptation environnementale chez cette espèce. Comme nous l'avons évoqué, plusieurs traits comme la diapause saisonnière ou l'acclimatation au froid sont des adaptations vraisemblablement centrales chez *Ae. albopictus*, assurant sa présence dans les environnements tempérés. Le polymorphisme de ces phénotypes peut être lié à une multitude de gènes ou séquences régulatrices impliquées dans diverses voies métaboliques. Il est aussi probable que d'autres traits, dont le polymorphisme n'a pas encore été mis en évidence, puissent jouer un rôle dans l'adaptation locale chez *Ae. albopictus*. Nous avons ainsi choisi de réaliser nos recherches sans *a priori*, en recherchant la signature de balayages sélectifs sur plusieurs dizaines de milliers de sites d'insertion polymorphes répartis le long du génome du moustique tigre.

Notre analyse a identifié 92 locus outliers, qui selon nos hypothèses, se trouveraient liés à une cible de la sélection naturelle. Notre premier réflexe a été de tenter de rechercher ces locus au sein des ressources génomiques disponibles, à savoir le transcriptiome d'*Ae. albopictus* et le génome annoté d'*Ae. aegypti*. Malheureusement, il s'est avéré impossible d'identifier avec certitude une position unique sur l'une ou l'autre de ces ressources. Si la présence d'un locus *outlier* parmi les régions transcrites du génome d'*Ae. albopictus* n'est pas impossible, il est très probable que les ET que nous suivons soient principalement insérés dans des régions non codantes ou non transcrites. D'autre part, le génome du moustique *Ae. aegypti* est définitivement trop éloigné du moustique tigre pour permettre l'identification de régions homologues : beaucoup des *outliers* peuvent être positionnés à plusieurs endroits et aucun n'obtient un alignement sur l'intégralité de sa séquence, même lorsque celle ci n'est que de quelques dizaines de paires de bases. En revanche, le

mapping des lectures non assemblées du génome d'*Ae. albopictus* sur les locus *outliers* (résultats non présentés) permet d'identifier ces locus comme appartenant bien au génome du moustique tigre.

Par ailleurs, la publication d'un premier assemblage du génome d'*Ae. albopictus* (Dritsou *et al.* 2015) devrait sous peu rendre disponibles les premières séquences sur lesquelles nous pourrons réitérer nos analyses. Cette publication est en fait la première de trois projets indépendants visant à assembler et annoter le génome du moustique tigre. Prochainement, la publication de la séquence d'une souche chinoise ainsi que de celle de la souche réunionnaise dont nous avons étudié le répétome devraient fournir de plus amples ressources afin de caractériser l'environnement génomique des locus *outliers*.

L'accès à ces régions du génome pourra permettre d'y confirmer l'action de la sélection naturelle. L'étude fine du polymorphisme génétique au voisinage des outliers devra révéler la signature caractéristique du balayage sélectif ; nous espérons alors pouvoir observer une baisse de la diversité génétique dabs ces régions par rapport à d'autres prises au hasard, voire des réductions locales du polymorphisme en leur sein. Il est par exemple possible d'amplifier puis de séquencer spécifiquement à forte couverture chez un grand nombre d'individus des régions plus larges au voisinage des *outliers* (méthodes type "capture de séquences", Peñalba *et al.* 2014). Ces analyses, nous permettrons ensuite d'identifier d'éventuels gènes ou séquences régulatrices pouvant être associés à l'adaptation climatique d'*Ae. albopictus*.

L'ultime étape, consistera alors à démontrer l'impact fonctionnel des variations génétiques mises en évidence entre Europe et Asie sur ces candidats. En particulier, il sera nécessaire de mesurer en conditions contrôlées leurs conséquences sur le phénotype et la valeur sélective des individus. Plusieurs approches sont pour cela envisageables. Si des gènes ou des séquences régulatrices peuvent être suspectées de jouer un rôle dans l'adaptation, leur mise sous silence (par exemple par ARN interférence) peut révéler la nature de leur influence sur le phénotype. Il faudra par ailleurs démontrer que ces différences fonctionnelles existent naturellement entre populations tempérées et tropicales ; ceci peut être réalisé en comparant entre elles le niveau d'expression d'un gène par PCR quantitative dans des conditions contrôlées. Par ailleurs, si l'analyse fine des régions candidates permet de prédire les conséquences fonctionnelles d'une variation génétique (présence d'un codon stop prématuré, insertions, délétions), les méthodes récentes d'édition du génome (en particulier à l'aide de CRISPR-Cas9), pourront permettre de provoquer spécifiquement la variation génétique considérée, en conservant par ailleurs le même fond génétique. Sans autre changements par ailleurs, l'impact de cette modification sur la *fitness* permettra de statuer sur le rôle adaptatif de la mutation.

Un autre résultat important de notre étude est qu'une large proportion de nos *outliers* se trouvent être des insertions à fortes fréquences en Europe. La grande majorité des insertions ségrégeant à faible fréquence, il semble très probable que seuls les auto-stops

génétiques au voisinage d'un allèle de type "présence" puisse entrainer une déviation significative des fréquences d'insertions entre les populations tempérées et tropicales. Dans le cas inverse, l'augmentation en fréquence au sein d'un groupe de populations d'un haplotype "absence" ne rendrait qu'encore plus rare les insertions portées par seulement quelques individus, et ne permettrait pas d'augmenter significativement la différenciation génétique au locus d'insertion. Sous cette hypothèse, nos résultats semblent indiquer que la plupart des récents évènements d'adaptation se sont produits chez des individus installés dans des milieux tempérés. Des populations adaptées à de tels environnements étant aussi établies au sein de l'aire d'origine, il est impossible sur la base de nos résultats d'établir si ces évènements ont eu lieu au cours de l'invasion, ou s'ils sont antérieurs à la colonisation et correspondent à des adaptations plus anciennes, déjà présentes par exemple au Japon ou en Chine. D'autre part, si ces *outliers* sont bien associés à l'adaptation aux environnements tempérés, serait-on capable de les retrouver chez d'autres populations invasives, notamment aux États-Unis ? Un tel cas de figure tendrait à conforter l'hypothèse d'une adaptation acquise au sein de l'aire d'origine.

Afin de la tester, le travail le plus immédiat consistera à tenter d'amplifier par PCR les *outliers* détectés lors du scan génomique chez de nouvelles populations de l'aire d'origine, tropicales et tempérées, ainsi que des populations invasives présentes sur d'autres continents. Nous possédons déjà plusieurs jeux d'amorces fonctionnels, permettant de réaliser cette expérience sur n'importe quel échantillon d'ADN génomique. L'une des limites de notre étude est d'avoir réalisé nos comparaisons avec seulement trois populations de l'aire d'origine, présentes elles-mêmes dans une région restreinte du Vietnam. Porretta *et al.* (2012) ont récemment confirmé à l'aide de marqueurs mitochondriaux hautement polymorphes, que la diversité génétique, au moins sur la partie continentale de l'Asie du sud-est, était particulièrement élevée. Cependant, la même étude révèle que cette diversité ne présente pas de structure géographique, et ce malgré un effort d'échantillonnage important du Bhoutan au Japon, en passant par la péninsule indochinoise et la Chine. La majorité des haplotypes est par ailleurs retrouvée en Thaïlande et au Vietnam, signe que cette région pourrait abriter une part importante de la diversité génétique au sein du berceau de l'espèce.
Les résultats de ce scan génomique bénéficieront donc de la comparaison avec de nouveaux échantillons.

Néanmoins, s'il était avéré que des adaptations s'étaient produites suite à l'invasion en Europe, nous pouvons essayer d'imaginer quels traits pourraient être sélectionnés. En particulier, nous pouvons nous demander si de tels évènements pourraient concerner la diapause photopériodique qui apparaît comme la caractéristique majeure du maintien annuel des populations tempérées. Ce phénotype étant décrit depuis longtemps au sein de l'aire d'origine (Hawley 1988), il serait parcimonieux de considérer que les populations invasives en Europe arborent ce trait du fait d'une ascendance avec des populations par

exemple japonaises ou chinoises. Cet argument, corroboré par les relevés des transports maritimes et quelques indices de génétique des populations, est notamment repris dans la littérature concernant l'invasion aux USA (Hawley *et al.* 1987 ; Kambhampati *et al.* 1991 ; Birungi et Munstermann 2002 ; Lounibos *et al.* 2003 ; Urbanski *et al.* 2012). Si tel était le cas, il semble tout de même possible que ce phénotype adaptatif puisse bénéficier de "réglages" lors de l'introduction dans un nouvel environnement. Les variations de la photopériode critique [1] (CPP) en fonction de la latitude, ont été comparées en conditions contrôlées à deux reprises, en 1988 et 2008, entre des populations invasives aux Etats-Unis et des populations japonaises supposées ancestrales (Urbanski *et al.* 2012). En 1988, les deux groupes de populations ne présentaient pas le même type de réponse (clines de pente différentes), les variations de CPP étant en particulier plus faibles et bien moins bien corrélées à la latitude aux USA ($r^2 = 0.35$) par rapport au Japon ($r^2 = 0.87$). Vingt ans plus tard, les mêmes populations américaines possédaient une très forte corrélation de leur CPP avec la latitude ($r^2 = 0.95$) et surtout la pente du cline mesurée n'était statistiquement plus différente de celle des populations japonaises. Enfin, bien que les pentes soient dorénavant identiques entre populations ancestrales et dérivées, les valeurs de CPP américaines étaient, à latitude égale, sensiblement plus basses que celles de leurs homologues japonaises. Les auteurs en ont donc conclu qu'une adaptation rapide aux USA avait conduit à la mise en place d'un nouveau cline pour la CPP. Certains facteurs, autres que la latitude, avaient alors pu conduire aux différences constatées entre les continents, ces "réglages" évoqués plus haut, qui pourraient être envisagés en Europe. D'un autre côté, sans preuve de l'origine des populations européennes, on peut se demander si des populations tropicales, introduites en Europe pourraient avoir été sélectionnées pour ce phénotype.

Une étude d'évolution expérimentale conduite sur une souche non-diapausante d'*Ae. albopictus*, n'a pas été en mesure de sélectionner pour ce trait (Craig 1993) ; cependant, Lounibos *et al.* (2003) pensent avoir identifié une souche brésilienne pouvant avoir acquis la diapause secondairement, ces auteurs considérant d'après la littérature que le Brésil aurait été colonisé par des individus non-diapausants et vraisemblablement d'origine tropicale. La diapause est un phénotype très répandu au sein du vivant, et implique une grande diversité de mécanismes conduisant à une dormance programmée (Schiesari et O'Connor 2013 ; Fenelon *et al.* 2014). Récemment Furness *et al.* (2015) ont montré que ce phénotype était apparu au moins 7 fois de manière indépendante chez les poissons du sous-ordre Aplocheiloidei (Killis). Chez les Culicidae, la diapause semble avoir elle aussi évoluée de manière parallèle étant donné la diversité des mécanismes observée entre des espèces parfois proches (Denlinger et Armbruster 2014). L'évolution de la diapause pourrait en fait s'appuyer sur l'existence de trait dit "pré-adaptatifs" comme la quiescence, qui

1. Nombre d'heures de lumière nécessaire pour induire la diapause à 50% des individus d'une population sensible

est en effet un phénotype répandu chez les moustiques du genre *Aedes*.

Comme nous le pointons du doigt dans la revue (chapitre 1), les routes d'invasions supposées ne disposent pas aujourd'hui d'un solide support par la génétique des populations. La plupart des hypothèses sont au contraire basées sur la conservation de traits phénotypiques (en particulier la diapause) et ont été pendant longtemps soutenus par l'analyse de marqueurs mitochondriaux peu polymorphes. Les analyses les plus récentes (Porretta *et al.* 2012 ; Zhong *et al.* 2013 ; Manni *et al.* 2015) révèlent en réalité que les populations étudiées abritent chacune une très forte diversité génétique, sans lien apparent avec la géographie, et ce, à l'échelle mondiale. Afin d'imaginer comment l'acquisition et la dispersion de mutations adaptatives peuvent avoir mené à l'adaptation du moustique tigre aux différents environnements dans lesquels il est présent aujourd'hui, une étude de génétique des populations à très large échelle, basée sur l'utilisation des marqueurs génétiques récemment développés (récent microsatellites et séquences mitochondriales Beebe *et al.* 2013 ; Zhong *et al.* 2013 ; Manni *et al.* 2015) et incorporant des méthodes permettant de tester plusieurs scénarios complexes (de type ABC) est aujourd'hui plus que nécessaire.

## Hypothèses sur l'importance des ET chez *Ae. albopictus*

Enfin, à la synthèse des travaux réalisés au cours de cette thèse, il est difficile de ne pas formuler des hypothèses quant à un éventuel rôle joué par les ET dans le succès invasif d'*Ae. albopictus*. Nous avons vu que l'activité des ET pouvait représenter une source non négligeable de variation d'autant plus que certaines familles d'ET peuvent être spécifiquement mobilisés en présence d'un stress environnemental ou génomique (Capy *et al.* 2000 ; Biémont et Vieira 2006 ; Fablet et Vieira 2011).

Différents exemples ont montré que la mobilisation des ET pouvait être à l'origine d'un certain nombre d'adaptation. Chez *D. melanogaster*, plusieurs dizaines d'insertions candidates ont ainsi pu être mise en évidence au sein des populations américaines, ayant facilité leur adaptation après leur migration hors d'Afrique (González *et al.* 2008, 2010). En particulier, les insertions des ET *Bari-Jheh* ou *pogo* ont pu par exemple être directement impliqués dans la régulation du stress oxydatif et dans la détoxication (Guio *et al.* 2014 ; Mateo *et al.* 2014). Nous pouvons aussi citer le cas évoqué précédemment des Drosophiles du Mount Carmel en Israël (Kim *et al.* 2014). L'accumulation d'ET dans certaines régions du génome ("îlots") sont également à l'origine d'un grand nombre de nouvelles variations génomiques (indels, duplication de gènes ou d'exons) associées à des variation phénotypiques adaptatives chez la fourmi invasive *Cordiocondyla obscurior* (Schrader *et al.* 2014). Ou encore, chez le moustique *Culex pipiens*, un ET de classe II inséré dans un gène co-

dant pour un récepteur protéique induit la résistance à une toxine de la bactérie *Bacillus sphaericus* utilisée comme insecticide larvaire (Darboux *et al.* 2007).

Notamment parce que leur mobilisation peut être délétère, les ET sont par ailleurs la cible de marques épigénétiques, visant à limiter leur activité (Slotkin et Martienssen 2007). Ces marques, qui peuvent être par exemple la méthylation des cytosines, ou encore la modification des histones, peuvent avoir une influence directe sur l'expression des gènes ou des séquences régulatrices au voisinage des ET, dont la mobilisation dans la lignée germinale sera alors à l'origine d'une variabilité mise à disposition de la sélection (Fablet et Vieira 2011 ; Jablonka 2013).

Mais parce que l'environnement lui même peut avoir une influence sur le polymorphisme épigénétique (Verhoeven *et al.* 2010 ; Zhang *et al.* 2013 ; Dixon *et al.* 2014 ; Fellous *et al.* 2015), les ET, de part leur présence dans des régions fonctionnelles, pourraient conduire à la régulation par l'environnement du phénotype. Les ET ne seraient alors plus seulement la source de variation adaptative, mais pourraient aussi faciliter l'évolution de la plasticité phénotypique. Chez l'abeille mellifère *Apis mellifera*, la méthylation *de novo* d'un certain nombre de gènes est par exemple à l'origine du développement de la larve vers une caste d'ouvrière ; de telles modifications épigénétiques pourrait de fait être à l'origine de la mise en place de réponses plastiques chez les insectes (Glastad *et al.* 2011).

Le moustique tigre est aujourd'hui un modèle biologique suscitant l'intérêt d'un grand nombre de chercheurs dans des disciplines aussi variées que l'épidémiologie, l'écologie, la génomique ou encore l'évolution des interactions symbiotiques. Toutes ces disciplines contribuent au développement des connaissances qui, en plus de servir les intérêts appliqués à la santé, ouvrent d'importantes perspectives de découvertes fondamentales. Ces travaux de thèse constituent une contribution aux connaissances empiriques concernant la génomique de l'adaptation, en particulier dans le contexte des invasions biologiques. Les travaux sur ce modèle singulier sont aussi à l'origine d'outils et de nouvelles ressources à disposition de la communauté, et permettent par ailleurs d'envisager de nouvelles pistes de recherche concernant l'évolution des génomes et de leurs constituants transposables au cours de l'évolution.

# Références bibliographiques


Albert, A. Y. K., Sawaya, S., Vines, T. H., Knecht, A. K., Miller, C. T., Summers, B. R., Balabhadra, S., Kingsley, D. M., and Schluter, D. *The genetics of adaptive shape shift in stickleback: pleiotropy and effect size.* **Evolution; international journal of organic evolution**, 62(1): 76–85, January 2008. doi: 10.1111/j.1558-5646.2007.00259.x.

Ali, S. R. and Rozeboom, L. E. *Comparative laboratory observations on selective mating of Aedes (Stegomyia) albopictus Skuse and A. (S.) polynesiensis Marks.* **Mosquito News**, 33 (1):23–28, 1973.

Allgood, D. W. and Yee, D. A. *Influence of resource levels, organic compounds and laboratory colonization on interspecific competition between the Asian tiger mosquito Aedes albopictus (Stegomyia albopicta) and the southern house mosquito Culex quinquefasciatus.* **Medical and veterinary entomology**, 28(3):273–86, September 2014. doi: 10.1111/mve.12047.

Alto, B. W., Bettinardi, D. J., and Ortiz, S. *Interspecific Larval Competition Differentially Impacts Adult Survival in Dengue Vectors.* **Journal of Medical Entomology**, 52(2):163–170, February 2015. doi: 10.1093/jme/tju062.

Antao, T. and Beaumont, M. a. *Mcheza: a workbench to detect selection using dominant markers.* **Bioinformatics (Oxford, England)**, 27(12):1717–8, June 2011. doi: 10.1093/bioinformatics/btr253.

Ayala, F. J., Serra, L., and Prevosti, A. *A grand experiment in evolution: the Drosophila subobscura colonization of the Americas.* **Genome**, 31(1):246–255, January 1989. doi: 10.1139/g89-042.

Bank, C., Ewing, G. B., Ferrer-Admettla, A., Foll, M., and Jensen, J. D. *Thinking too positive? Revisiting current methods of population genetic selection inference.* **Trends in Genetics**, 30(12):540–546, November 2014. doi: 10.1016/j.tig.2014.09.010.

Barnes, M. J., Lobo, N. F., Coulibaly, M. B., Sagnon, N. F., Costantini, C., and Besansky, N. J. *SINE insertion polymorphism on the X chromosome differentiates Anopheles gambiae molecular forms.* **Insect molecular biology**, 14(4):353–63, August 2005. doi: 10.1111/j.1365-2583.2005.00566.x.

Barrett, L. G., Thrall, P. H., Burdon, J. J., and Linde, C. C. *Life history determines genetic structure and evolutionary potential of host-parasite interactions.* **Trends in ecology & evolution**, 23(12):678–685, December 2008. doi: 10.1016/j.tree.2008.06.017.

Beaumont, M. a. *Adaptation and speciation: What can Fst tell us?* **Trends in Ecology and Evolution**, 20(8):435–440, 2005. doi: 10.1016/j.tree.2005.05.017.

Beaumont, M. a. and Balding, D. J. *Identifying adaptive genetic divergence among populations from genome scans.* **Molecular Ecology**, 13(4):969–980, April 2004. doi: 10.1111/j.1365-294X.2004.02125.x.

Beaumont, M. A. and Nichols, R. A. *Evaluating Loci for Use in the Genetic Analysis of Population Structure.* **Proceedings of the Royal Society London B**, 263(1377):1619–1626, December 1996. doi: 10.1098/rspb.1996.0237.

Beck, C. R., Garcia-Perez, J. L., Badge, R. M., and Moran, J. V. *LINE-1 elements in structural variation and disease.* **Annual review of genomics and human genetics**, 12:187–215, January 2011. doi: 10.1146/annurev-genom-082509-141802.

Beck, K. G., Zimmerman, K., Schardt, J. D., Stone, J., Lukens, R. R., Reichard, S., Randall, J., Cangelosi, A. A., Cooper, D., and Thompson, J. P. *Invasive Species Defined in a Policy Context: Recommendations from the Federal Invasive Species Advisory Committee.* **Invasive Plant Science and Management**, 1(4):414–421, October 2008. doi: 10.1614/IPSM-08-089.1.

Beebe, N. W., Ambrose, L., Hill, L. a., Davis, J. B., Hapgood, G., Cooper, R. D., Russell, R. C., Ritchie, S. a., Reimer, L. J., Lobo, N. F., Syafruddin, D., and van den Hurk, A. F. *Tracing the Tiger: Population Genetics Provides Valuable Insights into the Aedes (Stegomyia) albopictus Invasion of the Australasian Region.* **PLoS Neglected Tropical Diseases**, 7(8): e2361, August 2013. doi: 10.1371/journal.pntd.0002361.

Bellini, R., Albieri, A., Balestrino, F., Carrieri, M., Porretta, D., Urbanelli, S., Calvitti, M., Moretti, R., and Maini, S. *Dispersal and Survival of Aedes albopictus (Diptera: Culicidae) Males in Italian Urban Areas and Significance for Sterile Insect Technique Application.* **Journal of Medical Entomology**, 47(6):1082–1091, November 2010. doi: 10.1603/ME09154.

Biémont, C. and Vieira, C. *The influence of transposable elements on genome size.* **Journal de la Société de biologie**, 198(4):413–7, January 2004.

Biémont, C. and Vieira, C. *Genetics: junk DNA as an evolutionary force.* **Nature**, 443(7111): 521–4, October 2006. doi: 10.1038/443521a.

Bierne, N., Welch, J., Loire, E., Bonhomme, F., and David, P. *The coupling hypothesis: why genome scans may fail to map local adaptation genes.* **Molecular ecology**, 20(10):2044–72, May 2011. doi: 10.1111/j.1365-294X.2011.05080.x.

Birungi, J. and Munstermann, L. E. *Genetic Structure of Aedes albopictus (Diptera: Culicidae) Populations Based on Mitochondrial ND5 Sequences: Evidence for an Independent Invasion into Brazil and United States.* **Annals of the Entomological Society of America**, 95(1): 125–132, January 2002. doi: 10.1603/0013-8746(2002)095[0125:GSOAAD]2.0.CO;2.

Black, W. C. and Rai, K. S. *Genome evolution in mosquitoes: intraspecific and interspecific variation in repetitive DNA amounts and organization.* **Genetical Research**, 51(03):185, April 1988. doi: 10.1017/S0016672300024289.

Bock, D. G., Caseys, C., Cousens, R. D., Hahn, M. A., Heredia, S. M., Hübner, S., Turner, K. G., Whitney, K. D., and Rieseberg, L. H. *What we still don't know about invasion genetics.* **Molecular Ecology**, pages n/a–n/a, December 2015. doi: 10.1111/mec.13032.

Bonin, a., Ehrich, D., and Manel, S. *Statistical analysis of amplified fragment length polymorphism data: a toolbox for molecular ecologists and evolutionists.* **Molecular ecology**, 16(18): 3737–3758, September 2007. doi: 10.1111/j.1365-294X.2007.03435.x.

Bonin, A., Paris, M., Després, L., Tetreau, G., David, J.-P., and Kilian, A. *A MITE-based genotyping method to reveal hundreds of DNA polymorphisms in an animal genome after a few generations of artificial selection.* **BMC genomics**, 9:459, January 2008. doi: 10.1186/1471-2164-9-459.

Bonin, A., Paris, M., Tetreau, G., David, J.-P., and Després, L. *Candidate genes revealed by a genome scan for mosquito resistance to a bacterial insecticide: sequence and gene expression variations.* **BMC genomics**, 10(1):551, January 2009. doi: 10.1186/1471-2164-10-551.

Bonizzoni, M., Gasperi, G., Chen, X., and James, A. A. *The invasive mosquito species Aedes albopictus: current knowledge and future perspectives.* **Trends in parasitology**, 29(9):460–468, September 2013. doi: 10.1016/j.pt.2013.07.003.

Boulesteix, M., Simard, F., Antonio-Nkondjio, C., Awono-Ambene, H. P., Fontenille, D., and Biémont, C. *Insertion polymorphism of transposable elements and population structure of Anopheles gambiae M and S molecular forms in Cameroon.* **Molecular ecology**, 16(2):441–452, January 2007. doi: 10.1111/j.1365-294X.2006.03150.x.

Bourtzis, K., Dobson, S. L., Xi, Z., Rasgon, J. L., Calvitti, M., Moreira, L. A., Bossin, H. C., Moretti, R., Baton, L. A., Hughes, G. L., Mavingui, P., and Gilles, J. R. L. *Harnessing mosquito-Wolbachia symbiosis for vector and disease control.* **Acta tropica**, 132 Suppl:S150–63, April 2014. doi: 10.1016/j.actatropica.2013.11.004.

Boyer, S., Gilles, J., Merancienne, D., Lemperiere, G., and Fontenille, D. *Sexual performance of male mosquito Aedes albopictus.* **Medical and veterinary entomology**, 25(4):454–9, December 2011. doi: 10.1111/j.1365-2915.2011.00962.x.

Brady, O. J., Johansson, M. A., Guerra, C. A., Bhatt, S., Golding, N., Pigott, D. M., Delatte, H., Grech, M. G., Leisnham, P. T., Maciel-de Freitas, R., Styer, L. M., Smith, D. L., Scott, T. W., Gething, P. W., and Hay, S. I. *Modelling adult Aedes aegypti and Aedes albopictus survival at different temperatures in laboratory and field settings.* **Parasites & vectors**, 6(1):351, January 2013. doi: 10.1186/1756-3305-6-351.

Bridier-Nahmias, A., Tchalikian-Cosson, A., Baller, J. A., Menouni, R., Fayol, H., Flores, A., Saïb, A., Werner, M., Voytas, D. F., and Lesage, P. *An rna polymerase iii subunit determines sites of retrotransposon integration.* **Science**, 348(6234):585–588, 2015. doi: 10.1126/science.1259114.

Capy, P., Gasperi, G., Biémont, C., and Bazin, C. *Stress and transposable elements: co-evolution or useful parasites?* **Heredity**, 85(2):101–106, August 2000. doi: 10.1046/j.1365-2540.2000.00751.x.

Carnelossi, E. A. G., Lerat, E., Henri, H., Martinez, S., Carareto, C. M. A., and Vieira, C. *Specific activation of an I-like element in Drosophila interspecific hybrids.* **Genome biology and evolution**, 6(7):1806–17, July 2014. doi: 10.1093/gbe/evu141.

Casacuberta, E. and González, J. *The impact of transposable elements in environmental adaptation.* **Molecular ecology**, pages 1503–1517, January 2013. doi: 10.1111/mec.12170.

Catchen, J. M., Amores, A., Hohenlohe, P., Cresko, W., and Postlethwait, J. H. *Stacks: building and genotyping Loci de novo from short-read sequences.* **G3 (Bethesda, Md.)**, 1(3):171–182, August 2011. doi: 10.1534/g3.111.000240.

Chan, Y. F., Marks, M. E., Jones, F. C., Villarreal, G., Shapiro, M. D., Brady, S. D., Southwick, A. M., Absher, D. M., Grimwood, J., Schmutz, J., Myers, R. M., Petrov, D., Jónsson, B., Schluter, D., Bell, M. A., and Kingsley, D. M. *Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer.* **Science (New York, N.Y.)**, 327(5963):302–305, January 2010. doi: 10.1126/science.1182213.

Charlesworth, B., Morgan, M. T., and Charlesworth, D. *The effect of deleterious mutations on neutral molecular variation.* **Genetics**, 134(4):1289–1303, August 1993.

Chénais, B., Caruso, A., Hiard, S., and Casse, N. *The impact of transposable elements on eukaryotic genomes: from genome size increase to genetic adaptation to stressful environments.* **Gene**, 509(1):7–15, November 2012. doi: 10.1016/j.gene.2012.07.042.

Colautti, R. I. and Lau, J. A. *Contemporary evolution during invasion: evidence for differentiation, natural selection, and local adaptation.* **Molecular Ecology**, 24(9):1999–2017, May 2015. doi: 10.1111/mec.13162.

Craig, G. B. No Title The Diaspora of the Asian Tiger Mosquito. In McKnight, B. N., editor, *Biological pollution: the control and impact of invasive exotic species*, pages 101–120. Indiana Academy of Sciences, Indianapolis, 1993. ISBN 1883362008.

Danchin, E., Charmantier, A., Champagne, F. A., Mesoudi, A., Pujol, B., and Blanchet, S. *Beyond DNA: integrating inclusive inheritance into an extended theory of evolution.* **Nature reviews. Genetics**, 12(7):475–86, July 2011. doi: 10.1038/nrg3028.

Darboux, I., Charles, J.-F., Pauchet, Y., Warot, S., and Pauron, D. *Transposon-mediated resistance to Bacillus sphaericus in a field-evolved population of Culex pipiens (Diptera: Culicidae).* **Cellular microbiology**, 9(8):2022–9, August 2007. doi: 10.1111/j.1462-5822.2007.00934.x.

Darwin, C. R. *On the Origin of Species*. London, 1859.

Daub, J. T., Hofer, T., Cutivet, E., Dupanloup, I., Quintana-Murci, L., Robinson-Rechavi, M., and Excoffier, L. *Evidence for polygenic adaptation to pathogens in the human genome.* **Molecular biology and evolution**, 30(7):1544–58, July 2013. doi: 10.1093/molbev/mst080.

Daub, J. T., Dupanloup, I., Robinson-Rechavi, M., and Excoffier, L. *Inference of Evolutionary Forces Acting on Human Biological Pathways.* **Genome biology and evolution**, 7(6):1546–58, June 2015. doi: 10.1093/gbe/evv083.

de Lamballerie, X., Leroy, E., Charrel, R. N., Ttsetsarkin, K., Higgs, S., and Gould, E. A. *Chikungunya virus adapts to tiger mosquito via evolutionary convergence: a sign of things to come?* **Virology journal**, 5(1):33, January 2008. doi: 10.1186/1743-422X-5-33.

Delatte, H., Paupy, C., Dehecq, J., Thiria, J., Failloux, A., and Fontenille, D. *Aedes albopictus , vecteur des virus du chikungunya et de la dengue à la Réunion : biologie et contrôle.* **Parasite**, 15(1):3–13, March 2008. doi: 10.1051/parasite/2008151003.

Delatte, H., Bagny, L., Brengue, C., Bouetard, a., Paupy, C., and Fontenille, D. *The invaders: phylogeography of dengue and chikungunya viruses Aedes vectors, on the South West islands of the Indian Ocean.* **Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases**, 11(7):1769–1781, October 2011. doi: 10.1016/j.meegid.2011.07.016.

Denlinger, D. L. and Armbruster, P. A. *Mosquito diapause.* **Annual review of entomology**, 59:73–93, January 2014. doi: 10.1146/annurev-ento-011613-162023.

Dixon, G. B., Bay, L. K., and Matz, M. V. *Bimodal signatures of germline methylation are linked with gene expression plasticity in the coral Acropora millepora.* **BMC genomics**, 15(1):1109, January 2014. doi: 10.1186/1471-2164-15-1109.

Dlugosch, K. M. and Parker, I. M. *Founding events in species invasions: genetic variation, adaptive evolution, and the role of multiple introductions.* **Molecular ecology**, 17(1):431–449, January 2008. doi: 10.1111/j.1365-294X.2007.03538.x.

Dlugosch, K. M., Anderson, S. R., Braasch, J., Cang, F. A., and Gillette, H. D. *The devil is in the details: genetic variation in introduced populations and its contributions to invasion.* **Molecular Ecology**, pages n/a–n/a, April 2015. doi: 10.1111/mec.13183.

Dritsou, V., Topalis, P., Windbichler, N., Simoni, A., Hall, A., Lawson, D., Hinsley, M., Hughes, D., Napolioni, V., Crucianelli, F., Deligianni, E., Gasperi, G., Gomulski, L. M., Savini, G., Manni, M., Scolari, F., Malacrida, A. R., Arcà, B., Ribeiro, J. M., Lombardo, F., Saccone, G., Salvemini, M., Moretti, R., Aprea, G., Calvitti, M., Picciolini, M., Papathanos, P. A., Spaccapelo, R., Favia, G., Crisanti, A., and Louis, C. *A draft genome sequence of an invasive mosquito: an Italian Aedes albopictus.* **Pathogens and global health**, pages 207–220, September 2015. doi: 10.1179/2047773215Y.0000000031.

Duforet-Frebourg, N., Bazin, E., and Blum, M. G. B. *Genome scans for detecting footprints of local adaptation using a Bayesian factor model.* **Molecular biology and evolution**, 31(9): 2483–95, September 2014. doi: 10.1093/molbev/msu182.

Ellegren, H. *Microsatellites: simple sequences with complex evolution.* **Nature reviews. Genetics**, 5(6):435–445, 2004. doi: 10.1038/nrg1348.

Esnault, C., Boulesteix, M., Duchemin, J. B., Koffi, A. A., Chandre, F., Dabiré, R., Robert, V., Simard, F., Tripet, F., Donnelly, M. J., Fontenille, D., and Biémont, C. *High genetic differentiation between the M and S molecular forms of Anopheles gambiae in Africa.* **PloS one**, 3(4):e1968, January 2008. doi: 10.1371/journal.pone.0001968.

Estrada-Franco, J. G. and Graig, G. B. *Biology, disease relationships, and control of Aedes albopictus.* Pan American Health Organization, 1995.

Excoffier, L., Hofer, T., and Foll, M. *Detecting loci under selection in a hierarchically structured population.* **Heredity**, 103(4):285–98, October 2009. doi: 10.1038/hdy.2009.74.

Excoffier, L. and Ray, N. *Surfing during population expansions promotes genetic revolutions and structuration.* **Trends in ecology & evolution**, 23(7):347–351, July 2008. doi: 10.1016/j.tree.2008.04.004.

Fablet, M. and Vieira, C. *Evolvability, epigenetics and transposable elements.* **BioMolecular Concepts**, 2(5), January 2011. doi: 10.1515/BMC.2011.035.

Fay, J. C. and Wu, C. I. *Hitchhiking under positive Darwinian selection.* **Genetics**, 155(3): 1405–13, July 2000.

Fellous, A., Favrel, P., and Riviere, G. *Temperature influences histone methylation and mRNA expression of the Jmj-C histone-demethylase orthologues during the early development of the oyster Crassostrea gigas.* **Marine genomics**, 19:23–30, February 2015. doi: 10.1016/j.margen.2014.09.002.

Fenelon, J. C., Banerjee, A., and Murphy, B. D. *Embryonic diapause: development on hold.* **The International journal of developmental biology**, 58(2-4):163–74, January 2014. doi: 10.1387/ijdb.140074bm.

Ferrari, J. A. and Rai, K. S. *Phenotypic Correlates of Genome Size Variation in Aedes albopictus.* **Evolution**, 43(4):895, July 1989. doi: 10.2307/2409317.

Focks, D. A., Linda, S. B., Craig, G. B., Hawley, W. A., and Pumpuni, C. B. *Aedes albopictus (Diptera: Culicidae): A Statistical Model of the Role of Temperature, Photoperiod, and Geography in the Induction of Egg Diapause.* **Journal of Medical Entomology**, 31(2):278–286, March 1994. doi: 10.1093/jmedent/31.2.278.

Foll, M. and Gaggiotti, O. *A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective.* **Genetics**, 180(2):977–93, October 2008. doi: 10.1534/genetics.108.092221.

Foll, M., Gaggiotti, O. E., Daub, J. T., Vatsiou, A., and Excoffier, L. *Widespread signals of convergent adaptation to high altitude in Asia and america.* **American journal of human genetics**, 95(4):394–407, October 2014. doi: 10.1016/j.ajhg.2014.09.002.

Forattini, O. P. *Identificação de Aedes (Stegomyia) albopictus (Skuse) no Brasil.* **Revista de Saúde Pública**, 20(3):244–245, 1986.

Fouet, C., Gray, E., Besansky, N. J., and Costantini, C. *Adaptation to aridity in the malaria mosquito Anopheles gambiae: chromosomal inversion polymorphism and body size influence resistance to desiccation.* **PloS one**, 7(4):e34841, January 2012. doi: 10.1371/journal.pone.0034841.

Fraïsse, C., Belkhir, K., Welch, J. J., and Bierne, N. *Local interspecies introgression is the main cause of extreme levels of intraspecific differentiation in mussels.* **Molecular ecology**, July 2015. doi: 10.1111/mec.13299.

Furness, A. I., Reznick, D. N., Springer, M. S., and Meredith, R. W. *Convergent evolution of alternative developmental trajectories associated with diapause in African and South American killifish.* **Proceedings of the Royal Society B: Biological Sciences**, 282(1802):20142189–20142189, January 2015. doi: 10.1098/rspb.2014.2189.

Glastad, K. M., Hunt, B. G., Yi, S. V., and Goodisman, M. A. D. *DNA methylation in insects: on the brink of the epigenomic era.* **Insect molecular biology**, 20(5):553–65, October 2011. doi: 10.1111/j.1365-2583.2011.01092.x.

González, J., Lenkov, K., Lipatov, M., Macpherson, J. M., and Petrov, D. A. *High rate of recent transposable element-induced adaptation in Drosophila melanogaster.* **PLoS biology**, 6(10):e251, October 2008. doi: 10.1371/journal.pbio.0060251.

González, J., Karasov, T. L., Messer, P. W., and Petrov, D. A. *Genome-wide patterns of adaptation to temperate environments associated with transposable elements in Drosophila.* **PLoS genetics**, 6(4):e1000905, April 2010. doi: 10.1371/journal.pgen.1000905.

Goodier, J. L. and Kazazian, H. H. *Retrotransposons revisited: the restraint and rehabilitation of parasites.* **Cell**, 135(1):23–35, October 2008. doi: 10.1016/j.cell.2008.09.022.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., and Regev, A. *Full-length transcriptome assembly from RNA-Seq data without a reference genome.* **Nature biotechnology**, 29(7):644–52, July 2011. doi: 10.1038/nbt.1883.

Green, M. M. Mobile DNA Elements and Spontaneous Gene Mutation. In Lambert, M., McDonarld, J., and Weinstein, I., editors, *Eukaryotic Transposable Elements as Mutagenic Agents*, page 336. Banbury Report, 1988.

Greilhuber, J. *Intraspecific Variation in Genome Size: A Critical Reassessment.* **Annals of Botany**, 82(suppl_1):27–35, December 1998. doi: 10.1006/anbo.1998.0725.

Greilhuber, J. *Intraspecific variation in genome size in angiosperms: identifying its existence.* **Annals of botany**, 95(1):91–8, January 2005. doi: 10.1093/aob/mci004.

Grzebelus, D. *Transposon insertion polymorphism as a new source of molecular markers.* **Journal of fruit and ornamental plant research**, 14:21–29, 2006.

Gu, J., Orr, N., Park, S. D., Katz, L. M., Sulimova, G., MacHugh, D. E., and Hill, E. W. *A genome scan for positive selection in thoroughbred horses.* **PloS one**, 4(6):e5767, January 2009. doi: 10.1371/journal.pone.0005767.

Guio, L., Barrón, M. G., and González, J. *The transposable element Bari-Jheh mediates oxidative stress response in Drosophila.* **Molecular ecology**, 23(8):2020–2030, April 2014. doi: 10.1111/mec.12711.

Handley, L.-J., Estoup, A., Evans, D. M., Thomas, C. E., Lombaert, E., Facon, B., Aebi, A., and Roy, H. E. *Ecological genetics of invasive alien species.* **BioControl**, 56(4):409–428, August 2011. doi: 10.1007/s10526-011-9386-2.

Hanson, S. M. and Craig, G. B. *Cold Acclimation, Diapause, and Geographic Origin Affect Cold Hardiness in Eggs of Aedes albopictus (Diptera: Culicidae).* **Journal of Medical Entomology**, 31(2):192–201, March 1994. doi: 10.1093/jmedent/31.2.192.

Hawley, W., Reiter, P., Copeland, R., Pumpuni, C., and Craig, G. *Aedes albopictus in North America: probable introduction in used tires from northern Asia.* **Science**, 236(4805):1114–1116, May 1987. doi: 10.1126/science.3576225.

Hawley, W. A. *The biology of Aedes albopictus.* **Journal of the American Mosquito Control Association. Supplement**, 1:1–39, December 1988.

Hofer, T., Ray, N., Wegmann, D., and Excoffier, L. *Large allele frequency differences between human continental groups are more likely to have occurred by drift during range expansions than by selection.* **Annals of human genetics**, 73(1):95–108, January 2009. doi: 10.1111/j.1469-1809.2008.00489.x.

Hoffmann, A. A. and Weeks, A. R. *Climatic selection on genes and traits after a 100 year-old invasion: a critical look at the temperate-tropical clines in Drosophila melanogaster from eastern Australia.* **Genetica**, 129(2):133–47, February 2007. doi: 10.1007/s10709-006-9010-z.

Hohenlohe, P. a., Day, M. D., Amish, S. J., Miller, M. R., Kamps-Hughes, N., Boyer, M. C., Muhlfeld, C. C., Allendorf, F. W., Johnson, E. a., and Luikart, G. *Genomic patterns of introgression in rainbow and westslope cutthroat trout illuminated by overlapping paired-end RAD sequencing.* **Molecular Ecology**, pages n/a—-n/a, February 2013. doi: 10.1111/mec.12239.

Holsinger, K. E. Lecture note: Tajima's D, Fu's FS, Fay and Wu's H, and Zeng et al.'s E - http://darwin.eeb.uconn.edu/eeb348/lecturenotes/molevol-tajima.pdf, 2012.

Huang, X., Poelchau, M. F., and Armbruster, P. A. *Global Transcriptional Dynamics of Diapause Induction in Non-Blood-Fed and Blood-Fed Aedes albopictus.* **PLOS Neglected Tropical Diseases**, 9(4):e0003724, April 2015. doi: 10.1371/journal.pntd.0003724.

Iskow, R. C., McCabe, M. T., Mills, R. E., Torene, S., Pittard, W. S., Neuwald, A. F., Van Meir, E. G., Vertino, P. M., and Devine, S. E. *Natural mutagenesis of human genomes by endogenous retrotransposons.* **Cell**, 141(7):1253–1261, June 2010. doi: 10.1016/j.cell.2010.05.020.

Jablonka, E. *Epigenetic inheritance and plasticity: The responsive germline.* **Progress in Biophysics and Molecular Biology**, 111(2-3):99–107, April 2013. doi: 10.1016/j.pbiomolbio.2012.08.014.

Jaquiéry, J., Stoeckel, S., Nouhaud, P., Mieuzet, L., Mahéo, F., Legeai, F., Bernard, N., Bonvoisin, A., Vitalis, R., and Simon, J.-C. *Genome scans reveal candidate regions involved in the adaptation to host plant in the pea aphid complex.* **Molecular ecology**, 21(21):5251–64, November 2012. doi: 10.1111/mec.12048.

Kambhampati, S., Black, W. C., and Rai, K. S. *Geographic origin of the US and Brazilian Aedes albopictus inferred from allozyme analysis.* **Heredity**, 67 ( Pt 1)(September 1990): 85–93, August 1991.

Keller, R. P., Cadotte, M. W., and Sandiford, G. *Invasive Species in a Globalized World.* University of Chicago Press, 2014. ISBN 9780226166186. doi: 10.7208/chicago/9780226166216.001.0001.

Kesavaraju, B., Leisnham, P. T., Keane, S., Delisi, N., and Pozatti, R. *Interspecific Competition between Aedes albopictus and A. sierrensis: potential for Competitive Displacement in the Western United States.* **PloS one**, 9(2):e89698, January 2014. doi: 10.1371/journal.pone.0089698.

Kim, D. and Rossi, J. *RNAi mechanisms and applications.* **BioTechniques**, 44(5):613–6, April 2008. doi: 10.2144/000112792.

Kim, Y. B., Oh, J. H., McIver, L. J., Rashkovetsky, E., Michalak, K., Garner, H. R., Kang, L., Nevo, E., Korol, A. B., and Michalak, P. *Divergence of Drosophila melanogaster repeatomes in response to a sharp microclimate contrast in Evolution Canyon, Israel.* **Proceedings of the National Academy of Sciences**, 111(29):10630–10635, July 2014. doi: 10.1073/pnas.1410372111.

Kim, Y. *Allele frequency distribution under recurrent selective sweeps.* **Genetics**, 172(3):1967–78, March 2006. doi: 10.1534/genetics.105.048447.

Kirkpatrick, M. and Barrett, B. *Chromosome inversions, adaptive cassettes and the evolution of species' ranges.* **Molecular ecology**, January 2015. doi: 10.1111/mec.13074.

Klopfstein, S., Currat, M., and Excoffier, L. *The fate of mutations surfing on the wave of a range expansion.* **Molecular biology and evolution**, 23(3):482–90, March 2006. doi: 10.1093/molbev/msj057.

Koch, P., Platzer, M., and Downie, B. R. *RepARK-de novo creation of repeat libraries from whole-genome NGS reads.* **Nucleic acids research**, 42(9):e80, May 2014. doi: 10.1093/nar/gku210.

Kraemer, M. U. G., Sinka, M. E., Duda, K. A., Mylne, A., Shearer, F. M., Barker, C. M., Moore, C. G., Carvalho, R. G., Coelho, G. E., Van Bortel, W., Hendrickx, G., Schaffner, F., Elyazar, I. R., Teng, H.-J., Brady, O. J., Messina, J. P., Pigott, D. M., Scott, T. W., Smith, D. L., Wint, G. W., Golding, N., and Hay, S. I. *The global distribution of the arbovirus vectors Aedes aegypti and Ae. albopictus.* **eLife**, 4:e08347, June 2015. doi: 10.7554/eLife.08347.

Kulathinal, R. J., Bennett, S. M., Fitzpatrick, C. L., and Noor, M. A. F. *Fine-scale mapping of recombination rate in Drosophila refines its correlation to diversity and divergence.* **Proceedings of the National Academy of Sciences of the United States of America**, 105(29):10051–6, July 2008. doi: 10.1073/pnas.0801848105.

Kumar, A. and Rai, K. S. *Intraspecific variation in nuclear DNA content among world populations of a mosquito, Aedes albopictus (Skuse).* **TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik**, 79(6):748–52, July 1990. doi: 10.1007/BF00224239.

Lande, R. *Evolution of phenotypic plasticity in colonizing species.* **Molecular ecology**, January 2015. doi: 10.1111/mec.13037.

Lewontin, R. C. and Krakauer, J. *Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms.* **Genetics**, 74(1):175–195, May 1973.

Liew, C. and Curtis, C. F. *Horizontal and vertical dispersal of dengue vector mosquitoes, Aedes aegypti and Aedes albopictus, in Singapore.* **Medical and veterinary entomology**, 18(4): 351–60, December 2004. doi: 10.1111/j.0269-283X.2004.00517.x.

Lotterhos, K. E. and Whitlock, M. C. *Evaluation of demographic history and neutral parameterization on the performance of FST outlier tests.* **Molecular ecology**, pages 2178–2192, March 2014. doi: 10.1111/mec.12725.

Lounibos, L. P., Escher, R. L., and Lourenço-De-Oliveira, R. *Asymmetric Evolution of Photoperiodic Diapause in Temperate and Tropical Invasive Populations of <I>Aedes albopictus</I> (Diptera: Culicidae).* **Annals of the Entomological Society of America**, 96(4):512–518, July 2003. doi: 10.1603/0013-8746(2003)096[0512:AEOPDI]2.0.CO;2.

Lynch, M. and Conery, J. S. *The origins of genome complexity.* **Science (New York, N.Y.)**, 302(5649):1401–4, November 2003. doi: 10.1126/science.1089370.

Madon, M. B., Mulla, M. S., Shaw, M. W., Kluh, S., and Hazelrigg, J. E. *Introduction of Aedes albopictus (Skuse) in southern California and potential for its establishment.* **Journal of vector ecology : journal of the Society for Vector Ecology**, 27(1):149–54, June 2002.

Manni, M., Gomulski, L. M., Aketarawong, N., Tait, G., Scolari, F., Somboon, P., Guglielmino, C. R., Malacrida, A. R., and Gasperi, G. *Molecular markers for analyses of intraspecific genetic diversity in the Asian Tiger mosquito, Aedes albopictus.* **Parasites & vectors**, 8(1): 188, January 2015. doi: 10.1186/s13071-015-0794-5.

Marini, F., Caputo, B., Pombi, M., Tarsitani, G., and della Torre, A. *Study of Aedes albopictus dispersal in Rome, Italy, using sticky traps in mark-release-recapture experiments.* **Medical and veterinary entomology**, 24(4):361–8, December 2010. doi: 10.1111/j.1365-2915.2010.00898.x.

Mateo, L., Ullastres, A., and González, J. *A transposable element insertion confers xenobiotic resistance in Drosophila.* **PLoS genetics**, 10(8):e1004560, August 2014. doi: 10.1371/journal.pgen.1004560.

Maynard Smith, J. and Haigh, J. *The hitch-hiking effect of a favourable gene.* **Genetical Research**, 23(01):23, April 1974. doi: 10.1017/S0016672300014634.

McLain, D. K., Rai, K. S., and Fraser, M. J. *Intraspecific and interspecific variation in the sequence and abundance of highly repeated DNA among mosquitoes of the Aedes albopictus subgroup.* **Heredity**, 58(3):373–381, June 1987. doi: 10.1038/hdy.1987.65.

McVean, G. A. T., Myers, S. R., Hunt, S., Deloukas, P., Bentley, D. R., and Donnelly, P. *The fine-scale structure of recombination rate variation in the human genome.* **Science (New York, N.Y.)**, 304(5670):581–4, April 2004. doi: 10.1126/science.1092500.

Meier, K., Hansen, M. M., Bekkevold, D., Skaala, O., and Mensberg, K.-L. D. *An assessment of the spatial scale of local adaptation in brown trout (Salmo trutta L.): footprints of selection at microsatellite DNA loci.* **Heredity**, 106(3):488–99, March 2011. doi: 10.1038/hdy.2010.164.

Miller, M. R., Dunham, J. P., Amores, A., Cresko, W. A., and Johnson, E. A. *Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers.* **Genome research**, 17(2):240–8, February 2007. doi: 10.1101/gr.5681207.

Minard, G., Tran, F.-H., Dubost, A., Tran-Van, V., Mavingui, P., and Moro, C. V. *Pyrosequencing 16S rRNA genes of bacteria associated with wild tiger mosquito Aedes albopictus: a pilot study.* **Frontiers in cellular and infection microbiology**, 4:59, January 2014. doi: 10.3389/fcimb.2014.00059.

Modolo, L. and Lerat, E. *Identification and analysis of transposable elements in genomic sequences.* **Genome analysis: Current Procedures and Applications**, pages 165–181, 2014.

Monden, Y. and Tahara, M. *Plant transposable elements and their application to genetic analysis via high-throughput sequencing platform.* **The Horticulture Journal**, advpub, 2015. doi: 10.2503/hortj.MI-IR02.

Monden, Y., Yamamoto, A., Shindo, A., and Tahara, M. *Efficient DNA fingerprinting based on the targeted sequencing of active retrotransposon insertion sites using a bench-top high-throughput sequencing platform.* **DNA research : an international journal for rapid publication of reports on genes and genomes**, 21(5):491–8, October 2014. doi: 10.1093/dnares/dsu015.

Mori, A., Oda, T., and Wada, Y. *Studies on the egg diapause and overwintering of Aedes albopictus in Nagasaki.* **Tropical Medicine**, 23(2):79–90, 1981.

Müller, G. C., Xue, R.-D., and Schlein, Y. *Differential attraction of Aedes albopictus in the field to flowers, fruits and honeydew.* **Acta tropica**, 118(1):45–9, April 2011. doi: 10.1016/j.actatropica.2011.01.009.

Mutebi, J.-P., Black, W. C., Bosio, C. F., Sweeney, W. P., and Craig, G. B. *Linkage Map for the Asian Tiger Mosquito [Aedes (Stegomyia) albopictus] Based on SSCP Analysis of RAPD Markers.* **Journal of Heredity**, 88(6):489–494, November 1997. doi: 10.1093/oxfordjournals.jhered.a023142.

Nachman, M. *Variation in recombination rate across the genome: evidence and implications.* **Current Opinion in Genetics & Development**, 12(6):657–663, December 2002. doi: 10.1016/S0959-437X(02)00358-1.

Nei, M. and Maruyama, T. *Letters to the editors: Lewontin-Krakauer test for neutral genes.* **Genetics**, 80(2):395, June 1975.

Niebylski, M. and Craig, G. *Dispersal and survival of Aedes albopictus at a sap tire yard in Missouri.* **Journal of the American Mosquito Control Association**, 10(3):339–343, September 1994.

Nielsen, R. *Molecular signatures of natural selection.* **Annual review of genetics**, 39:197–218, January 2005. doi: 10.1146/annurev.genet.39.073003.112420.

Novák, P., Neumann, P., and Macas, J. *Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data.* **BMC bioinformatics**, 11:378, January 2010. doi: 10.1186/1471-2105-11-378.

Oleksyk, T. K., Smith, M. W., and O'Brien, S. J. *Genome-wide scans for footprints of natural selection.* **Philosophical transactions of the Royal Society of London. Series B, Biological sciences**, 365(1537):185–205, January 2010. doi: 10.1098/rstb.2009.0219.

Oliva, C. F., Damiens, D., Vreysen, M. J. B., Lemperière, G., and Gilles, J. *Reproductive strategies of Aedes albopictus (Diptera: Culicidae) and implications for the sterile insect technique.* **PloS one**, 8(11):e78884, January 2013. doi: 10.1371/journal.pone.0078884.

Orr, H. A. *The genetic theory of adaptation: a brief history.* **Nature reviews. Genetics**, 6 (2):119–27, February 2005. doi: 10.1038/nrg1523.

Paupy, C., Delatte, H., Bagny, L., Corbel, V., and Fontenille, D. *Aedes albopictus, an arbovirus vector: from the darkness to the light.* **Microbes and infection / Institut Pasteur**, 11 (14-15):1177–85, December 2009. doi: 10.1016/j.micinf.2009.05.005.

Peñalba, J. V., Smith, L. L., Tonione, M. A., Sass, C., Hykin, S. M., Skipwith, P. L., McGuire, J. A., Bowie, R. C. K., and Moritz, C. *Sequence capture using PCR-generated probes: a cost-effective method of targeted high-throughput sequencing for nonmodel organisms.* **Molecular ecology resources**, 14(5):1000–10, September 2014. doi: 10.1111/1755-0998.12249.

Peischl, S., Dupanloup, I., Kirkpatrick, M., and Excoffier, L. *On the accumulation of deleterious mutations during range expansions.* **Molecular ecology**, 22(24):5972–82, December 2013. doi: 10.1111/mec.12524.

Peischl, S. and Excoffier, L. *Expansion load: recessive mutations and the role of standing genetic variation.* **Molecular ecology**, March 2015. doi: 10.1111/mec.13154.

Pérez-Figueroa, A., García-Pereira, M. J., Saura, M., Rolán-Alvarez, E., and Caballero, A. *Comparing three different methods to detect selective loci using dominant markers.* **Journal of evolutionary biology**, 23(10):2267–76, October 2010. doi: 10.1111/j.1420-9101.2010.02093.x.

Poelchau, M. F., Reynolds, J. a., Denlinger, D. L., Elsik, C. G., and Armbruster, P. a. *A de novo transcriptome of the Asian tiger mosquito, Aedes albopictus, to identify candidate transcripts for diapause preparation.* **BMC genomics**, 12(1):619, January 2011. doi: 10.1186/1471-2164-12-619.

Poelchau, M. F., Reynolds, J. a., Denlinger, D. L., Elsik, C. G., and Armbruster, P. a. *Transcriptome sequencing as a platform to elucidate molecular components of the diapause response in the Asian tiger mosquito, Aedes albopictus.* **Physiological entomology**, 38(2):173–181, April 2013a. doi: 10.1111/phen.12016.

Poelchau, M. F., Reynolds, J. a., Elsik, C. G., Denlinger, D. L., and Armbruster, P. a. *Deep sequencing reveals complex mechanisms of diapause preparation in the invasive mosquito, Aedes albopictus.* **Proceedings. Biological sciences / The Royal Society**, 280(1759): 20130143, January 2013b. doi: 10.1098/rspb.2013.0143.

Poelchau, M. F., Reynolds, J. a., Elsik, C. G., Denlinger, D. L., and Armbruster, P. a. *RNA-Seq reveals early distinctions and late convergence of gene expression between diapause and quiescence in the Asian tiger mosquito, Aedes albopictus.* **The Journal of experimental biology**, 216(Pt 21):4082–4090, November 2013c. doi: 10.1242/jeb.089508.

Porretta, D., Mastrantonio, V., Bellini, R., Somboon, P., and Urbanelli, S. *Glacial history of a modern invader: phylogeography and species distribution modelling of the Asian tiger mosquito Aedes albopictus.* **PloS one**, 7(9):e44515, January 2012. doi: 10.1371/journal.pone.0044515.

Prevosti, A., Ribo, G., Serra, L., Aguade, M., Balana, J., Monclus, M., and Mestres, F. *Colonization of America by Drosophila subobscura: Experiment in natural populations that supports the adaptive role of chromosomal-inversion polymorphism.* **Proceedings of the National Academy of Sciences**, 85(15):5597–5600, August 1988. doi: 10.1073/pnas.85.15.5597.

Rai, K. S. and Black, W. C. *Mosquito genomes: structure, organization, and evolution.* **Advances in genetics**, 41:1–33, January 1999.

Ran, F. A., Hsu, P. D., Lin, C.-Y., Gootenberg, J. S., Konermann, S., Trevino, A. E., Scott, D. A., Inoue, A., Matoba, S., Zhang, Y., and Zhang, F. *Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity.* **Cell**, 154(6):1380–9, September 2013. doi: 10.1016/j.cell.2013.08.021.

Rao, P. N. and Rai, K. S. *Inter and intraspecific variation in nuclear DNA content in Aedes mosquitoes.* **Heredity**, 59(2):253–258, October 1987. doi: 10.1038/hdy.1987.120.

Robertson, A. *Gene frequency distribution as a test for selective neutrality.* **Genetics**, 81(4): 775–785, December 1975.

Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z. P., Richter, D. J., Schaffner, S. F., Gabriel, S. B., Platko, J. V., Patterson, N. J., McDonald, G. J., Ackerman, H. C., Campbell, S. J., Altshuler, D., Cooper, R., Kwiatkowski, D., Ward, R., and Lander, E. S. *Detecting recent positive selection in the human genome from haplotype structure.* **Nature**, 419(6909): 832–7, October 2002. doi: 10.1038/nature01140.

Sabot, F., Picault, N., El-Baidouri, M., Llauro, C., Chaparro, C., Piegu, B., Roulin, A., Guiderdoni, E., Delabastide, M., McCombie, R., and Panaud, O. *Transpositional landscape of the rice genome revealed by paired-end mapping of high-throughput re-sequencing data.* **The Plant Journal**, 66(2):241–246, 2011. doi: 10.1111/j.1365-313X.2011.04492.x.

Savolainen, O., Lascoux, M., and Merilä, J. *Ecological genomics of local adaptation.* **Nature reviews. Genetics**, 14(11):807–20, November 2013. doi: 10.1038/nrg3522.

Schaffner, F., Van Bortel, W., and Coosemans, M. *First record of Aedes (Stegomyia) albopictus in Belgium.* **Journal of the American Mosquito Control Association**, 20(2):201–203, 2004.

Schiesari, L. and O'Connor, M. B. *Diapause: delaying the developmental clock in response to a changing environment.* **Current topics in developmental biology**, 105:213–46, January 2013. doi: 10.1016/B978-0-12-396968-2.00008-7.

Schrader, L., Kim, J. W., Ence, D., Zimin, A., Klein, A., Wyschetzki, K., Weichselgartner, T., Kemena, C., Stökl, J., Schultner, E., Wurm, Y., Smith, C. D., Yandell, M., Heinze, J., Gadau, J., and Oettler, J. *Transposable element islands facilitate adaptation to novel environments in an invasive species.* **Nature communications**, 5:5495, January 2014. doi: 10.1038/ncomms6495.

Sexton, J. P., Hangartner, S. B., and Hoffmann, A. A. *Genetic isolation by environment or distance: which pattern of gene flow is most common?* **Evolution; international journal of organic evolution**, 68(1):1–15, January 2014. doi: 10.1111/evo.12258.

Sivan, A., Shriram, A. N., Sunish, I. P., and Vidhya, P. T. *Host-feeding pattern of Aedes aegypti and Aedes albopictus (Diptera: Culicidae) in heterogeneous landscapes of South Andaman, Andaman and Nicobar Islands, India.* **Parasitology research**, 114(9):3539–3546, July 2015. doi: 10.1007/s00436-015-4634-5.

Slotkin, R. K. and Martienssen, R. *Transposable elements and the epigenetic regulation of the genome.* **Nature reviews. Genetics**, 8(4):272–85, April 2007. doi: 10.1038/nrg2072.

Sota, T. and Mogi, M. *Interspecific variation in desiccation survival time of Aedes (Stegomyia) mosquito eggs is correlated with habitat and egg size.* **Oecologia**, 90(3):353–358, June 1992. doi: 10.1007/BF00317691.

Sprenger, D. and Wuithiranyagool, T. *The discovery and distribution of Aedes albopictus in Harris County, Texas.* **Journal of the American Mosquito Control Association**, 2(2): 217–219, June 1986.

Stapley, J., Santure, A. W., and Dennis, S. R. *Transposable elements as agents of rapid adaptation may explain the genetic paradox of invasive species.* **Molecular ecology**, 24(9):2241–52, May 2015. doi: 10.1111/mec.13089.

Stephan, W. *Genetic hitchhiking versus background selection: the controversy and its implications.* **Philosophical transactions of the Royal Society of London. Series B, Biological sciences**, 365(1544):1245–53, April 2010. doi: 10.1098/rstb.2009.0278.

Stephan, W. *Signatures of positive selection: from selective sweeps at individual loci to subtle allele frequency changes in polygenic adaptation.* **Molecular Ecology**, pages n/a–n/a, June 2015. doi: 10.1111/mec.13288.

Tajima, F. *Statistical method for testing the neutral mutation hypothesis by DNA polymorphism.* **Genetics**, 123(3):585–95, November 1989.

Templeton, A. *Population Genetics and Microevolutionary Theory - Alan R. Templeton.* John Wiley & Sons, Inc, Hoboken, New Jersey, 2006. ISBN 978-0-471-40951-9.

Tiffin, P. and Ross-Ibarra, J. *Advances and limits of using population genetics to understand local adaptation.* **Trends in Ecology & Evolution**, 29(12):673–680, November 2014. doi: 10.1016/j.tree.2014.10.004.

Tsuda, Y., Maekawa, Y., Ogawa, K., Itokawa, K., Komagata, O., Sasaki, T., Isawa, H., Tomita, T., and Sawabe, K. *Biting density and distribution of Aedes albopictus during the September 2014 outbreak of dengue fever in Yoyogi Park and the vicinity in Tokyo Metropolis, Japan.* **Japanese journal of infectious diseases**, March 2015. doi: 10.7883/yoken.JJID.2014.576.

Urbanski, J., Mogi, M., O'Donnell, D., DeCotiis, M., Toma, T., and Armbruster, P. *Rapid adaptive evolution of photoperiodic response during invasion and range expansion across a climatic gradient.* **The American naturalist**, 179(4):490–500, April 2012. doi: 10.1086/664709.

Vela, D., Fontdevila, A., Vieira, C., and García Guerreiro, M. P. *A genome-wide survey of genetic instability by transposition in Drosophila hybrids.* **PloS one**, 9(2):e88992, January 2014. doi: 10.1371/journal.pone.0088992.

Verhoeven, K. J. F., Jansen, J. J., Van Dijk, P. J., and Biere, A. *Stress-induced DNA methylation changes and their heritability in asexual dandelions.* **New Phytologist**, 185(4):1108–1118, March 2010. doi: 10.1111/j.1469-8137.2009.03121.x.

Vitalis, R., Dawson, K., and Boursot, P. *Interpretation of Variation Across Marker Loci as Evidence of Selection.* **Genetics**, 158(4):1811–1823, August 2001.

Vitek, C. J. and Livdahl, T. *Hatch Plasticity in Response to Varied Inundation Frequency in <I>Aedes albopictus</I>.* **Journal of Medical Entomology**, 46(4):766–771, July 2009. doi: 10.1603/033.046.0406.

Voight, B. F., Kudaravalli, S., Wen, X., and Pritchard, J. K. *A map of recent positive selection in the human genome.* **PLoS biology**, 4(3):e72, March 2006. doi: 10.1371/journal.pbio.0040072.

Vos, P., Hogers, R., Bleeker, M., Reijans, M., van de Lee, T., Hornes, M., Friters, A., Pot, J., Paleman, J., Kuiper, M., and Zabeau, M. *AFLP: a new technique for DNA fingerprinting.* **Nucleic Acids Research**, 23(21):4407–4414, February 1995. doi: 10.1093/nar/23.21.4407.

Wanlapakorn, N., Thongmee, T., Linsuwanon, P., Chattakul, P., Vongpunsawad, S., Payungporn, S., and Poovorawan, Y. *Chikungunya outbreak in Bueng Kan Province, Thailand, 2013.* **Emerging infectious diseases**, 20(8):1404–6, August 2014. doi: 10.3201/eid2008.140481.

Wasserberg, G., Bailes, N., Davis, C., and Yeoman, K. *Hump-shaped density-dependent regulation of mosquito oviposition site-selection by conspecific immature stages: theory, field test with Aedes albopictus, and a meta-analysis.* **PloS one**, 9(3):e92658, January 2014. doi: 10.1371/journal.pone.0092658.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., Paux, E., SanMiguel, P., and Schulman, A. H. *A unified classification system for eukaryotic transposable elements.* **Nature reviews. Genetics**, 8 (12):973–82, December 2007. doi: 10.1038/nrg2165.

Williams, G. C. *Pleiotropy, Natural Selection, and the Evolution of Senescence.* **Evolution**, 11 (4):398, December 1957. doi: 10.2307/2406060.

Williges, E., Faraji, A., and Gaugler, R. *Vertical Oviposition Preferences of the Asian Tiger Mosquito, Aedes albopictus, In Temperate North America.* **Journal of the American Mosquito Control Association**, 30(3):169–74, September 2014. doi: 10.2987/14-6409R.1.

Witherspoon, D. J., Xing, J., Zhang, Y., Watkins, W. S., Batzer, M. a., and Jorde, L. B. *Mobile element scanning (ME-Scan) by targeted high-throughput sequencing.* **BMC genomics**, 11: 410, January 2010. doi: 10.1186/1471-2164-11-410.

Yee, D. A. and Skiff, J. F. *Interspecific Competition of a New Invasive Mosquito, <I>Culex coronator</I>, and Two Container Mosquitoes, Aedes albopictus and Cx. quinquefasciatus (Diptera: Culicidae), Across Different Detritus Environments.* **Journal of Medical Entomology**, 51(1):89–96, January 2014. doi: 10.1603/ME13182.

Yee, D. A., Juliano, S. A., and Vamosi, S. M. *Seasonal Photoperiods Alter Developmental Time and Mass of an Invasive Mosquito, Aedes albopictus (Diptera: Culicidae), Across Its North-South Range in the United States.* **Journal of Medical Entomology**, 49(4):825–832, July 2012. doi: 10.1603/ME11132.

Zhang, Y.-Y., Fischer, M., Colot, V., and Bossdorf, O. *Epigenetic variation creates potential for evolution of plant phenotypic plasticity.* **The New phytologist**, 197(1):314–22, January 2013. doi: 10.1111/nph.12010.

Zhivotovsky, L. A. *Estimating population structure in diploids with multilocus dominant DNA markers.* **Molecular Ecology**, 8(6):907–913, June 1999. doi: 10.1046/j.1365-294x.1999.00620.x.

Zhong, D., Lo, E., Hu, R., Metzger, M. E., Cummings, R., Bonizzoni, M., Fujioka, K. K., Sorvillo, T. E., Kluh, S., Healy, S. P., Fredregill, C., Kramer, V. L., Chen, X., and Yan, G. *Genetic analysis of invasive Aedes albopictus populations in Los Angeles County, California and its potential public health impact.* **PloS one**, 8(7):e68586, January 2013. doi: 10.1371/journal.pone.0068586.

Zytnicki, M., Akhunov, E., and Quesneville, H. *Tedna: a Transposable Element De Novo Assembler.* **Bioinformatics (Oxford, England)**, pages 1–2, June 2014. doi: 10.1093/bioinformatics/btu365.

# Annexes

# Annexe 1 : Minard et al., 2015

Les populations étudiées lors du scan génomique sont issues d'un échantillonnage réalisé sur le terrain par Guillaume Minard, Claire Valiente Moro et Patrick Mavingui, avec qui nous avons collaboré tout au long de cette thèse. L'article présenté ici correspond à une partie du travail de thèse réalisé par Guillaume, visant à comparer la composition du microbiote de l'intestin moyen chez ces mêmes populations. La diversité du microbiote à été étudiée en relation avec l'environnement et la structure génétique des hôtes, analysée à l'aide de marqueurs microsatellites et mitochondriaux.

Ma participation à cet article concerne les aspects de génétique des populations, en tant que conseil dans le choix et l'interprétation des analyses réalisées par Guillaume.

L'un des résultats principaux est l'observation d'une réduction significative de la diversité microbienne au sein des populations invasives. Celle-ci est aussi accompagnée d'une légère baisse de la diversité génétique aux locus microsatellites. Les populations Vietnamiennes et Françaises restent cependant très peu différenciées génétiquement et aucune corrélation n'a pu être trouvée entre la composition du microbiote et la structure génétique. Ces résultats laissent supposer que malgré des effets de dérive liés à l'introduction, l'environnement conserve un rôle important dans l'établissement de la diversité microbienne symbiotique.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 22 September 2015
doi: 10.3389/fmicb.2015.00970

# French invasive Asian tiger mosquito populations harbor reduced bacterial microbiota and genetic diversity compared to Vietnamese autochthonous relatives

*G. Minard[1], F. H. Tran[1], Van Tran Van[1], C. Goubert[2], C. Bellet[3], G. Lambert[4], Khanh Ly Huynh Kim[5], Trang Huynh Thi Thuy[5], P. Mavingui[1,6] and C. Valiente Moro[1]**

[1] Ecologie Microbienne, UMR Centre National de la Recherche Scientifique 5557, USC INRA 1364, VetAgro Sup, FR41 BioEnvironment and Health, Université Claude Bernard Lyon 1, Villeurbanne, France, [2] Laboratoire de Biométrie et Biologie Evolutive, UMR 5558, CNRS, INRIA, VetAgro Sup, Villeurbanne, France, [3] Entente Interdépartementale Rhône-Alpes pour la Démoustication, Chindrieux, France, [4] Entente Interdépartementale de Démoustication du Littoral Méditerranéen, Montpellier, France, [5] Department of Medical Entomology and Zoonotics, Pasteur Institute in Ho Chi Minh City, Vietnam, [6] Université de La Réunion, UMR PIMIT, INSERM U1187, CNRS 9192, IRD 249, Plateforme Technologique CYROI, Saint-Denis, France

OPEN ACCESS

**Edited by:**
*Joerg Graf,*
*University of Connecticut, USA*

**Reviewed by:**
*David William Waite,*
*University of Auckland, New Zealand*
*Brian Weiss,*
*Yale University, USA*

***Correspondence:**
*C. Valiente Moro,*
*Ecologie Microbienne, Université Claude Bernard Lyon 1, Bat. André Lwoff, 10 Rue Raphaël Dubois, 69100 Villeurbanne, France*
*claire.valiente-moro@univ-lyon1.fr*

**Specialty section:**
*This article was submitted to Microbial Symbioses, a section of the journal Frontiers in Microbiology*

**Received:** *24 July 2015*
**Accepted:** *01 September 2015*
**Published:** *22 September 2015*

**Citation:**
*Minard G, Tran FH, Van VT, Goubert C, Bellet C, Lambert G, Kim KLH, Thuy THT, Mavingui P and Valiente Moro C (2015) French invasive Asian tiger mosquito populations harbor reduced bacterial microbiota and genetic diversity compared to Vietnamese autochthonous relatives. Front. Microbiol. 6:970. doi: 10.3389/fmicb.2015.00970*

The Asian tiger mosquito *Aedes albopictus* is one of the most significant pathogen vectors of the twenty-first century. Originating from Asia, it has invaded a wide range of eco-climatic regions worldwide. The insect-associated microbiota is now recognized to play a significant role in host biology. While genetic diversity bottlenecks are known to result from biological invasions, the resulting shifts in host-associated microbiota diversity has not been thoroughly investigated. To address this subject, we compared four autochthonous *Ae. albopictus* populations in Vietnam, the native area of *Ae. albopictus*, and three populations recently introduced to Metropolitan France, with the aim of documenting whether these populations display differences in host genotype and bacterial microbiota. Population-level genetic diversity (microsatellite markers and COI haplotype) and bacterial diversity (16S rDNA metabarcoding) were compared between field-caught mosquitoes. Bacterial microbiota from the whole insect bodies were largely dominated by *Wolbachia pipientis*. Targeted analysis of the gut microbiota revealed a greater bacterial diversity in which a fraction was common between French and Vietnamese populations. The genus *Dysgonomonas* was the most prevalent and abundant across all studied populations. Overall genetic diversities of both hosts and bacterial microbiota were significantly reduced in recently established populations of France compared to the autochthonous populations of Vietnam. These results open up many important avenues of investigation in order to link the process of geographical invasion to shifts in commensal and symbiotic microbiome communities, as such shifts may have dramatic impacts on the biology and/or vector competence of invading hematophagous insects.

**Keywords:** ***Aedes albopictus*, *Dysgonomonas*, holobiont, microbiota, microsatellite, phylogeography, *Wolbachia***

## Introduction

Mosquitoes are considered by the World Health Organization to be the most medically important disease vectors. The Asian tiger mosquito (*Aedes albopictus*) is of major concern as it is known to be able to carry 26 arboviruses including Dengue and Chikungunya (Paupy et al., 2009). Furthermore, *Ae. albopictus* is considered as one of the most geographically invasive species. It has rapidly spread from its native area of South and East Asia to reach various eco-climatic regions in America, Africa, Oceania and Europe (Bonizzoni et al., 2013). The worldwide trades in secondhand tires and lucky bamboo, both of which often contain standing water making them ideal places for mosquito eggs and larvae, have been key factors in *Ae. albopictus* transportation. Once established in a new region, the tiger mosquito easily adapts and persists in a wide range of habitats, even in temperate climates mainly due to its aptitude to enter into a state of dormancy or "diapause" (Urbanski et al., 2010). Undoubtedly, the intrinsic capacities of the mosquito populations largely play an important role in their ecological plasticity. However, this assumption remains surprising as according to the "paradox of invasion," recent introductions often imply a burden for the genetic structure of newly introduced populations (reviewed by Handley et al., 2011).

A comprehensive understanding of insect population genetics now requires an integrative approach considering microorganisms as a key component of the system. According to the hologenome theory, Metazoan organisms should no longer be considered as individuals, but rather as holobionts consisting of the host plus all of its associated microorganisms (Zilber-Rosenberg and Rosenberg, 2008). It is the holobiont and its associated hologenome that can be considered as a unit of selection which is impacted by variation, selection, drift and evolution (reviewed by Rosenberg and Zilber-Rosenberg, 2014). Insect holobionts are also difficult to decipher as they may include a range of host-symbiont relationships ranging from parasitism to mutualism (Toft and Andersson, 2010). Numerous studies have demonstrated the contribution of the microbiota to the biology of the host (Blottière et al., 2013; Douglas, 2014). Some mutualistic symbionts favor ecological adaptations in insects (Douglas, 2011, 2014) by playing key roles in extended phenotypes such as growth, nutrition, reproduction, protection against pathogens and tolerance to environmental stresses (Buchner, 1965; Dillon and Dillon, 2004; Moran et al., 2008; Moya et al., 2008). Moreover, the host genotype may also influence in return symbiont communities (Ochman et al., 2010; Muegge et al., 2011). It is thus, important to consider the genetic basis of bacterial microbiota selection.

It is striking that relatively few studies have focused on bacterial microbiota associated with invasive arthropods. Many invasive species harbor genetic modifications caused by founder effects or genetic drift, but the effect of such changes on their microbiota are only just beginning to be documented (Meusnier et al., 2001; Zurel et al., 2011; Ye et al., 2014). As symbiotic microorganisms can change more rapidly and by more diverse processes than the host organism itself they could influence the adaptation and evolution of the holobiont.

Here, we aimed to document whether populations from two representative areas colonized by the Asian tiger mosquito displayed changes in their host genotypes and associated bacterial microbiota. To that end, we sampled mosquitoes from field populations in Vietnam, a country located in the South East Asia where *Ae. albopictus* originated indicating ancient colonization, and in Metropolitan France, which was more recently invaded by this species. Mitochondrial and nuclear genotypes of mosquitoes and bacterial diversity were compared within and between populations.

## Materials and Methods

### Sampling Areas and Mosquito Collection

*Ae. albopictus* specimens were sampled in Metropolitan France and Vietnam. In addition to their contrasted climate and ecology, these two countries were chosen as Vietnam is located in the South East Asia, the region where *Ae. albopictus* originated indicating ancient colonization, whereas Metropolitan France is a newly invaded zone. Sampling in Metropolitan France was performed between August and September 2012 at Saint-Priest (SP), Portes-Lès-Valence (PLV), and Nice (NC). NC is one of the first invaded sites in France since 2004 (Medlock et al., 2012), whereas PLV and SP were colonized in 2011 and 2012, respectively (data obtained from French health organization INVS). Mosquito sampling in Vietnam was performed during October 2012 at Hồ Chí Minh City (HCM), Bình Du'o'ng (BD), Vung Tàu City (VT), and Bù Gia Mâp (BGM) (**Figure 1**). All sites were urban or suburban, except BGM located in a protected forest national park. Sampling sites were at least 18 km from each other to avoid sampling populations originated from the same breeding site. Consequently, we assume that individuals collected from the same sites belong to the same population as they share breeding sites, and a total of seven independent populations (three from Metropolitan France and four from Vietnam) were obtained and analyzed. Live adult females were caught with nets or BG-Traps and then identified using morphological characteristics (Rueda, 2004). For some individuals, identification was confirmed by COI barcode sequencing (see below). To control confounding effects from nutritional factors, only females that could be seen to contain no blood upon magnified observation of the gut contents were retained for analysis. Mosquitoes were stored in 100% ethanol at −80°C until used.

### Quantification of *Wolbachia w*AlbA and *w*AlbB Strains

*Wolbachia w*AlbA and *w*AlbB were quantified in whole bodies of 10 mosquitoes from each site (**Table S1**). Their densities were measured in triplicate by qPCR amplification of the *Wolbachia wsp* gene and normalized with the *Ae. albopictus actin* gene as described (Zouache et al., 2009). Standard curves were drawn on DNA plasmid pQuantAlb which contains *wsp* genes of *Wolbachia pipientis w*AlbA and *w*AlbB as well as *actin* gene of *Ae. albopictus* (Tortosa et al., 2008). Correlations between the two strains were calculated with R software using the Pearson's correlation.

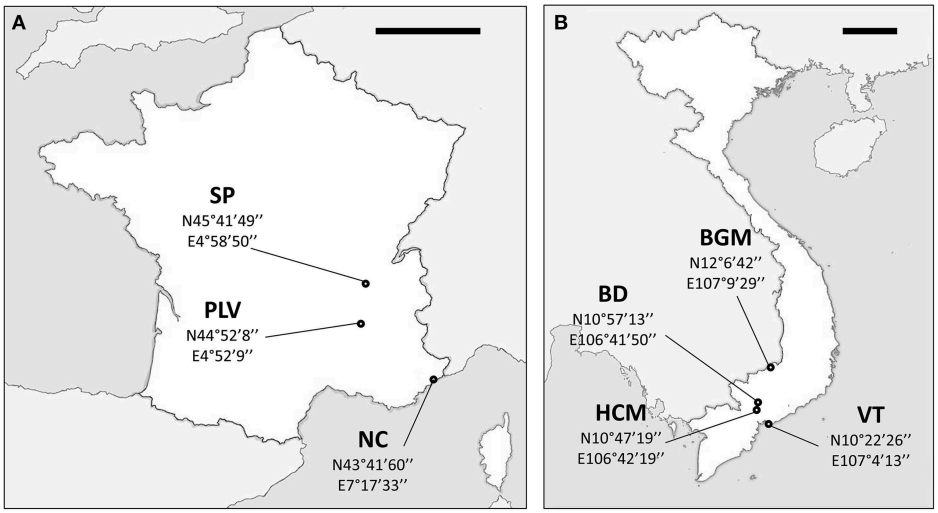



**FIGURE 1 | Maps of sampling sites.** Sampling sites and GPS coordinates (*World Geodetic System 1984*) in France **(A)** and Vietnam **(B)**. NC, Nice; PLV, Porte-lès-Valence; SP, Saint Priest; VT, Vung Tàu City; HCM, Hồ Chí Minh City; BD, Bình Du'o'ng, BGM, Bù Gia Mập. Scale bars, 200 km.

## Sample Preparation, Miseq Sequencing, Quality Trimming and Diversity Analysis of Sequences

Previous work (Minard et al., 2014) showed that whole insect body was inappropriate for in-depth analysis of bacterial microbiota of *Ae. albopictus* by NGS due to the overrepresentation of *Wolbachia* sequences. Here we used midguts that are known to be a key organ in the metabolism and immunity of mosquitoes as well as the first point of entry for transmitted viruses (Clements, 1992; Jupatanakul et al., 2014; Kenney and Brault, 2014b). In addition this organ was shown to harbor a moderate density of *Wolbachia* in *Ae. albopictus* (Zouache et al., 2009).

Prior to dissection for midgut recovery from 32 individuals (**Table S1**), female specimens were surface-disinfected with 70% ethanol and rinsed with sterile water as previously described (Minard et al., 2013). All dissection steps were performed under a sterile laminar flow hood in a containment environment. Mosquitoes were dissected in sterile 1 × phosphate buffered saline solution (Life Technologies, NY, USA). For each mosquito, the midgut was separated from the rest of the body. Midguts were then individually crushed with 1-mm diameter beads in ATL lysis buffer (Qiagen, Hilden, Germany) containing 20 mg.ml$^{-1}$ lysozyme (Euromedex, Strasbourg, France) using a Bioblock Scientific MM 2000 mill (Retsch, Eragny sur Oise, France). DNA was then extracted with Qiagen DNeasy Blood and Tissue kit (Qiagen, Hilden, Germany) following the manufacturer's recommendations for both Gram negative and Gram positive bacteria. Assuming that the remaining mosquito bodies (hereafter referred to as carcasses) should be dominated by *Wolbachia*, they were pooled per population (**Table S1**) and DNA extracted as above, and then used as positive controls. Finally, as negative controls to evaluate potential contamination, DNA extraction was carried out without any biological matrix and four independent eluates were concentrated and pooled.

Hypervariable V5-V6 *rrs* regions were amplified in triplicate for each DNA sample with 30 ng of DNA and modified primers 784F (5′-AGGATTAGATACCCTGGTA-3′) and 1061R (5′-CRRCACGAGCTGACGAC-3′) as described (Andersson et al., 2008) with modifications. Briefly, primers containing a 8-bp multiplex barcode and Illumina adapters were used for PCR amplifications with 1.75 U of Expand High Fidelity Enzyme Mix (Roche, Basel, Switzerland), 1 × Expand High Fidelity Buffer (Roche, Basel, Switzerland), 0.06 mg mL$^{-1}$ of T4 gene 32 protein (New England Biolabs, Evry, France), 0.06 mg mL$^{-1}$ of bovine serum albumin (New England Biolabs, Evry, France), 40 μM of dNTP mix, 200 nM of each primer (Life Technologies, Saint Aubin, France). Amplifications were carried out on Biorad C1000 thermal cycler (Biorad, CA, USA) with 5 min at 95°C, followed by 40 cycles at 95°C for 40 s, 54.2°C for 1 min, 72°C for 30 s, with a final extension step of 7 min at 72°C. Forty PCR amplification cycles were necessary to generate an optimal amount of amplicons for Miseq sequencing. The three PCR product replicates from each sample were pooled, purified with Agencourt AMPure XP PCR Purification kit (Beckman Coulter, Villepinte, France), and quantified using the Quant-iT Picogreen dsDNA Assay Kit (Life Technologies, NY, USA).

A total of 40 amplicon libraries were constructed: 32 for individual midguts, 7 for carcasses and 1 for the negative control. Sequencing of each library was performed on the Illumina MiSeq platform (2 × 250-bp paired-end reads) by ProfileXpert—ViroScan 3D (Lyon, France). All FastQ files were deposited at EMBL European Nucleotide Archive (https://www.ebi.ac.uk/ena) under the project accession number PRJEB6896. A total of 9, 222, 165 reads were obtained, paired-end reads were joined with PandaSeq (Masella et al., 2012), trimmed and aligned on the SILVA database release 115 using standard filtering tools in the MOTHUR pipeline (Schloss et al., 2009). Two errors were allowed in primer sequences, read sizes were filtered to

be 200–350 bp in length with no ambiguous bases. Chimeras were detected and removed with Perseus implemented in Mothur package. Based on the analysis of clustered sequence rates from 92 to 99% similarity, OTUs were re-adjusted to 97% similarity using a median neighbor algorithm. Sequences were classified according to the SILVA database release 115 at 80% minimum bootstrap using a naïve Bayesian classifier (Wang et al., 2007). OTUs were also kept if there were at least represented by more than one sequence overall samples. Furthermore, OTUs were removed from further analyses if they were detected in the negative control sample and their relative abundance was not at least 10 times greater than that observed in the negative control. This additional quality control criterion allows us to qualify and correct for low concentration contaminants of experimental origin. Richness, α and β diversity indices were calculated using a subsample of the same read number for each sample. Diversity analyses, hierarchical analysis of molecular variance (AMOVA), Non-Metric Multidimensional Scaling ordination and heatmap representations were performed with R software (R Development Core Team, 2009) using ade4 and vegan packages (Dray and Dufour, 2007; Oksanen et al., 2013). To highlight possible country-associated OTUs, extended errors bars were computed and classified according to Welch modified *t*-test significance ($p < 0.05$) using STAMP software 2.0.9 (Parks and Beiko, 2010).

### Mitochondrial Gene Amplification and Haplotyping

To maximize the quality and quantity of DNA obtained, our previously optimized extraction protocol for individual whole mosquito was used (Minard et al., 2014). A total of 85 individuals were analyzed ranging from 9 to 20 individuals per sampling site (**Table S1**). The 597-bp region of mtDNA cytochrome c oxidase subunit I (*COI*) gene was amplified with CI-J-1632 (5′-TGATCAAATTTATAAT-3′) and CI-N-2191 (5′-GGTAAAATTAAAATATAAACTTC-3′) primers using 45 ng of DNA matrix as described (Raharimalala et al., 2012). To analyze haplotypes, *COI* sequences were aligned with Seaview 4, then diversity and nucleotide composition were calculated with DnaSP (Librado and Rozas, 2009). AMOVA statistical analyses were performed with Arlequin 3.5 × (Excoffier and Lischer, 2010). Sequences of the different *COI* haplotypes were deposited on Genbank (https://www.ncbi.nlm.nih.gov/genbank) under the accession numbers LM999972-LM999977.

### Microsatellite Processing and Genotyping

A total of 199 individuals were genotyped ranging from 22 to 30 individuals per sampling site (**Table S1**). Amplifications were done with 10 ng of DNA extracted from each individual and master mix from Qiagen Type-it Microsatellite PCR Kit following the manufacturer's recommendations (Qiagen, Hilden, Germany). Amplifications were performed on Biorad C1000 thermal cycler (Biorad, CA, USA) with an optimal protocol for each microsatellite to minimize unspecific artifacts. A set of 11 microsatellite markers previously described (Chambers et al., 1986; Porretta et al., 2006; Beebe et al., 2013) were used; namely AealbA9, AealbB51, AealbB52, AEDC, Alb-di6, Alb-tri3, Alb-tri18, Alb-tri25, Alb-tri41, Alb-tri45, and Alb-tri6 (**Table S2**). For AeablA9, AealbB51, and AealbB52 microsatellites, the program consisted of 5 min at 94°C, followed by 35 amplification cycles at 94°C for 5 min, 52°C (AealbA9) or 50°C (AealbB51, Aealb0B52) for 30 s and 72°C for 45 s, with a final step of 30 min at 72°C. AEDC microsatellite sequences were amplified with 5 min at 94°C followed by 30 cycles of 45 s at 94°C, 1 min 30 s at 56°C and 45 s at 72°C, with a final step of 30 min at 60°C. Alb-di6, Alb-tri3, Alb-tri18, Alb-tri25, Alb-tri41, Alb-tri45, and Alb-tri6 were amplified as described by Beebe et al. (2013). PCR products were diluted (between 1/60 and 1/100 depending on the relative sensitivity of markers) then 1 μL was mixed with 13.8 μL of ultrapure Hi-Di-formamide TM and 0.2 μL of size marker (MRL 500) and loaded on an ABI Prism 3730XL Genetic Analyzer automated sequencer (Life Technologies, NY, USA). Microsatellites were scored manually with Genemapper 3.0 (LifeTechnologies, NY, USA). Null alleles were evaluated with FreeNA (Chapuis and Estoup, 2007). Diversity indices, linkage disequilibrium, Factorial Correspondence Analysis and hierarchical AMOVA analyses were computed with Genetix 4.05, Fstat 2.9.3.2 and Arlequin 3.5x softwares (Excoffier and Lischer, 2010). The Bayesian structure of populations was evaluated using Structure 2.3.4 (Pritchard et al., 2000) with 100 000 "burn-in" steps followed by 500 000 iterations. Runs from 1 to 8 potential groups (K) were processed with 20 replicates (**Figure S1**). An admixture model was used with a location prior factor. As recommended for datasets with possible null alleles, a dominant allele option was set in the model. The best fit K-value was chosen with the Evanno method implemented in STRUCTURE HARVESTER (Evanno et al., 2005; Earl and VonHoldt, 2012) and the 20 replicates were averaged with CLUMPP (Jakobsson and Rosenberg, 2007). Finally, a population structure barplot was drawn with DISTRUCT (Rosenberg, 2004). Comparisons between Fst distances and bacterial microbiota Bray-Curtis distances were performed with a Mantel test. Rarefied genetic richness (Ar) and diversity (Hs) were correlated with Shannon bacterial diversity using Spearman's rank correlation. As populations which experienced a recent reduction of their effective size can develop a heterozygosity excess at neutral loci, this parameter was tested using BOTTLENECK software (Cornuet and Luikart, 1996).

## Results

### *Wolbachia* *w*AlbA and *w*AlbB Strains are Abundant and Positively Correlated with Each Other

Our aim was to study the bacterial microbiota in autochthonous and invasive tiger mosquito populations. In a previous study, we demonstrated that *Wolbachia* is the predominant bacterial species in *Ae. albopictus* from Madagascar when using whole body genomic DNA, constituting up to 99% of high throughput sequences recovered (Minard et al., 2014). Here, we first tested whether the two *Wolbachia* strains *w*AlbA and *w*AlbB were also present and dominant in mosquitoes sampled from autochthonous populations at four sites (HCM, BD, VT, BGM) in Vietnam and from invasive populations at three sites (SP, PLV,

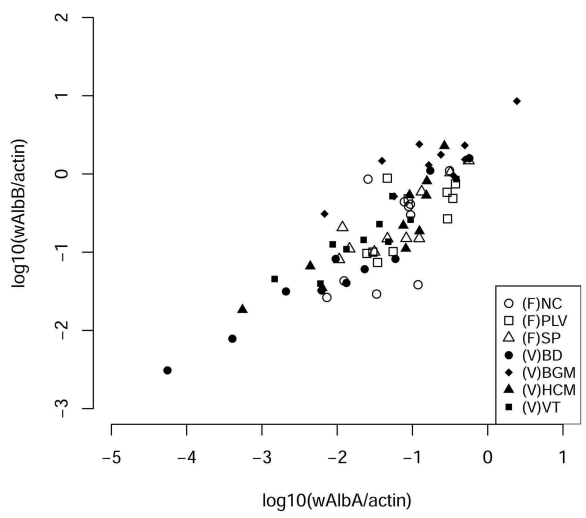



**FIGURE 2 | Correlation densities of *Wolbachia pipientis*.** The number of each bacterial strain per cells was evaluated by quantification of *wsp* genes from each *Wolbachia* *w*AlbA and *w*AlbB strains normalized to the number of host actin gene copies (Pearson's correlation $R^2 = 0.84$, $p < 2.2 \times 10^{-16}$). NC, Nice; PLV, Porte-lès-Valence; SP, Saint Priest; VT, Vung Tàu City; HCM, Hồ Chí Minh City; BD, Bình Du'o'ng, BGM, Bù Gia Mập.

NC) in France (**Figure 1**). *w*AlbA and *w*AlbB were detected in all individuals of the seven populations. The lowest densities of *w*AlbA and *w*AlbB strains detected were $5.25 \times 10^{-5}$ wsp.actin$^{-1}$ and $3.09 \times 10^{-3}$ wsp.actin$^{-1}$ copies respectively in mosquito samples from BD and the highest densities of 2.44 wsp.actin$^{-1}$ and 8.53 wsp.actin$^{-1}$ copies respectively from the BGM population (**Figure 2**). A significant positive correlation (Pearson's correlation $R^2 = 0.84$, $p < 2.2 \times 10^{-16}$) between *w*AlbA and *w*AlbB strains was found among all the populations tested (**Figure 2**). To check if *Wolbachia* sequences were overrepresented in bacterial microbiota sequences when applying NGS methods to the whole body, the V5-V6 *rrs* amplicons were generated from mosquito carcass pools as indicated in material and methods, and sequenced by Miseq technology. Results confirmed a dominance of *Wolbachia* OTUs (**Figure 3**) which account for 28% (for HCM) to 91% (for SP) of the sequence dataset, reinforcing the rationale for our choice to avoid using the whole insect body for bacterial community analysis in *Ae. albopictus*.

## Midgut Bacterial Community Structure in Vietnamese Autochthonous Populations Compared to French Invasive Ones

The insect gut is a key organ in insect physiology and immunity. Moreover, previous studies have demonstrated that this organ harbored low concentration of *Wolbachia* in *Ae. albopictus* adults (Zouache et al., 2009), opening up the possibility to extend the depth of analysis of the gut-associated microbial community. For this purpose, V5-V6 *rrs* amplicons from 32 individual midgut samples (from 3 to 5 individuals per sampling site) were sequenced with MiSeq technology. Analysis of a negative control showed the presence of bacterial sequences that probably derived from contamination during laboratory sample handling (**Table S6**). However, the diversity of this control was dissimilar from those of all mosquito samples (Bray-Curtis dissimilarity > 68.6%). For subsequent analysis of sequences associated with mosquito samples, OTUs potentially originating from laboratory contamination were trimmed from the whole dataset. Based on this analysis, a total of 2,088 OTUs were identified in all the midgut samples (between 306 and 1,272 OTUs per sample), with a total of 68 OTUs exceeding 1% in abundance. These OTU numbers were consistent with those previously obtained by high throughput sequencing of midguts from various mosquito species (Osei-Poku et al., 2012). The genus *Dysgonomonas* was the most prevalent and abundant OTU retrieved from the midgut samples (**Figure S3**), although its abundance varied from 3% (HCM8) to 72% (SP7) between samples (**Figure 3**). AMOVA analysis and ordinations were performed to detect the degree of differentiation at various hierarchical levels. No significant variation was observed between sites, indicating a low variability between individuals belonging to a given population. In contrast, variation between countries was found to be significant for the β-diversity measure, explaining a large part of the variation for Bray-Curtis dissimilarities (AMOVA, 22.79%, $p < 10^{-4}$) (**Table 1**, **Figure 4A**). Similar structures were obtained with Unifrac phylogeny based β-diversity distances (**Table S3**). Consequently, further comparisons of populations were performed at the country level. When compared with the four populations from Vietnam, consisting of a total of 14 individuals, the three populations from France composed of 18 individuals harbored less diverse and more homogeneous bacterial microbiota, indicated by lower values for the Chao 1 richness estimator (Mann–Whitney Wilcoxon, $p = 0.002$), Shannon-Weather α-diversity (Mann–Whitney Wilcoxon, $p < 10^{-3}$) and variance of abundance-weighted β-diversity (**Figure S2**). However, Bray-Curtis (Beta-dispersion, $p = 0.33$) distances were not significantly different between populations of the two countries.

Overall, 15 OTUs were shown to be significantly associated with Vietnamese populations (Welch corrected *t*-test, $p < 0.05$) (**Figure 4B**), including various members of *Shingomonadaceae* family (*Sphingobium*, *Novosphingobium*, *Sphingomonas*). Only three OTUs assigned to *Dysgonomonas* and *Aeromonas* genera and *Enterobacteriaceae* family were shown to be significantly associated with French populations. The OTU overlap was calculated after merging and subsampling sequences at the country level (**Figure 4C**). The resulting Venn diagram showed 21% of shared OTUs ($n = 309$), and 34% ($n = 503$), and 45% ($n = 656$) specific of France and Vietnam, respectively. Shared OTUs counted for 85% of the overall sequences.

## Low Mitochondrial DNA Variation among *Ae. albopictus* Populations from France and Vietnam

The genetic makeup of *Ae. albopictus* hosts was evaluated by barcoding based on mitochondrial *COI* gene sequences. A total of 6 haplotype variants were found in individuals from all

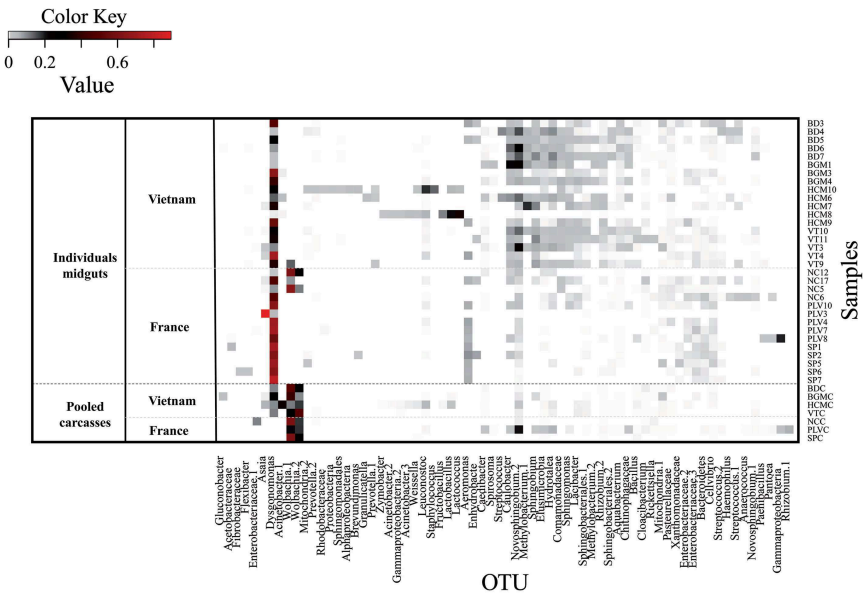



**FIGURE 3 | *Aedes albopictus* microbiota bacterial diversity.** The heatmap represents the relative abundance of 3% OTUs in each sample. The Color Key Value refers to the proportion of an OTU in a single sample. The OTUs were represented if their proportion was =0.01 in at least one sample. They were named by their taxonomic assignment according to Bayesian classification. When more than one OTU get the same assignation, they were differentiated with numbers. Samples are classified according to their own populations. Individual midgut samples from Vietnam were collected in Binh Du'o'ng (BD 3, 4, 5, 6, and 7), Bù Gia Mâp (BGM 1, 3, 4), Hồ Chí Minh City (HCM 6, 7, 8, 9, 10), Vung Tàu City (VT 3, 4, 9, 10, 11). Individual midgut samples from France were collected in Nice (NC 5, 6, 12, 17), Portes-Lès-Valence (PLV 3, 4, 7, 8, 10), and Saint Priest (SP 1, 2, 5, 6, 7). The pooled carcasses of the following individuals were also analyzed for both Vietnamese (BDC, BGMC, HCMC, VTC) and French populations (NCC, PLVC, SPC); the last "C" added to each name indicates carcass.

**TABLE 1 | AMOVA analysis.**

| | Haplotypes | | | Microsatellites | | | β-Diversity (Bray-Curtis) | | |
|---|---|---|---|---|---|---|---|---|---|
| | *df** | Variance (%) | *p* | *df** | Variance (%) | *p* | *df** | Variance (%) | *p* |
| Among countries | 1 | 54.3 | 0.02 | 1 | 2.61 | 0.2 | 1 | 22.79 | $<10^{-4}$ |
| Among populations/Countries | 5 | 20.54 | $<10^{-3}$ | 5 | 12.61 | $<10^{-3}$ | 5 | 7.14 | 0.48 |
| Within populations | 78 | 25.15 | | 391 | 84.78 | | 26 | 70.07 | |

**df, degree of freedom.*

populations. The genetic heterogeneity was more pronounced between countries (AMOVA 47.4%, $p < 10^{-3}$) than between sites (AMOVA 19.79%, $p < 10^{-3}$) (**Table 1**). The two major subclades were distinguishable by only one mutation. One subclade included 52 haplotypes (H_1) and the other 28 haplotypes (H_3) (**Table S4**). H_1 was mostly associated with the populations from Vietnam whereas H_3 was more associated with those from France. However, a mix of both haplotypes was also found in the populations at the PLV site in France (12/20 for H_1 and 7/20 for H_3) and at the BD site in Vietnam (16/17 for H_1 and 1/17 for H_3) (**Figure S4**).

## Evidence for Genetic Reduction in *Ae. albopictus* Populations Invasive to France

The mosquito nuclear genomic variation was characterized further by genotyping 199 individuals with 11 microsatellite markers. The overall number of alleles per locus varied from 6 (AealbB52) to 30 (Alb-tri 18). Rarefied allele richness of populations from each site was 4.56 for SP, 4.57 for PLV, 5.91 for NC, 6.50 for HCM, 7.47 for VT, 8.25 for BGM, and 8.62 for BD. To test for hypothetical population bottlenecks or expansions, the allelic richness (*Ar*) and the heterozygosity (He) at each genomic locus were analyzed for each sampling site. For both these analyses of variation, values for populations from France were significantly lower than those from Vietnam (Mann–Whitney, $p = 0.0003$ and 0.0003, respectively). The bottleneck analysis did not show any significant heterozygosity excess under a two-phase model or a single stepwise model (**Table S5**). No significant linkage disequilibrium was found between any pair of loci. All populations had a positive inbreeding index ($F_{IS}$) between 0.12 (for PLV) and 0.196 (for HCM) reflecting an excess of homozygotes (**Table 2**). Moreover, there were significant scores for the presence of null alleles. However, $F_{ST}$ values of the entire mosquito sample were 0.142 ($CI_{95\%}$ 0.058–0.269) with

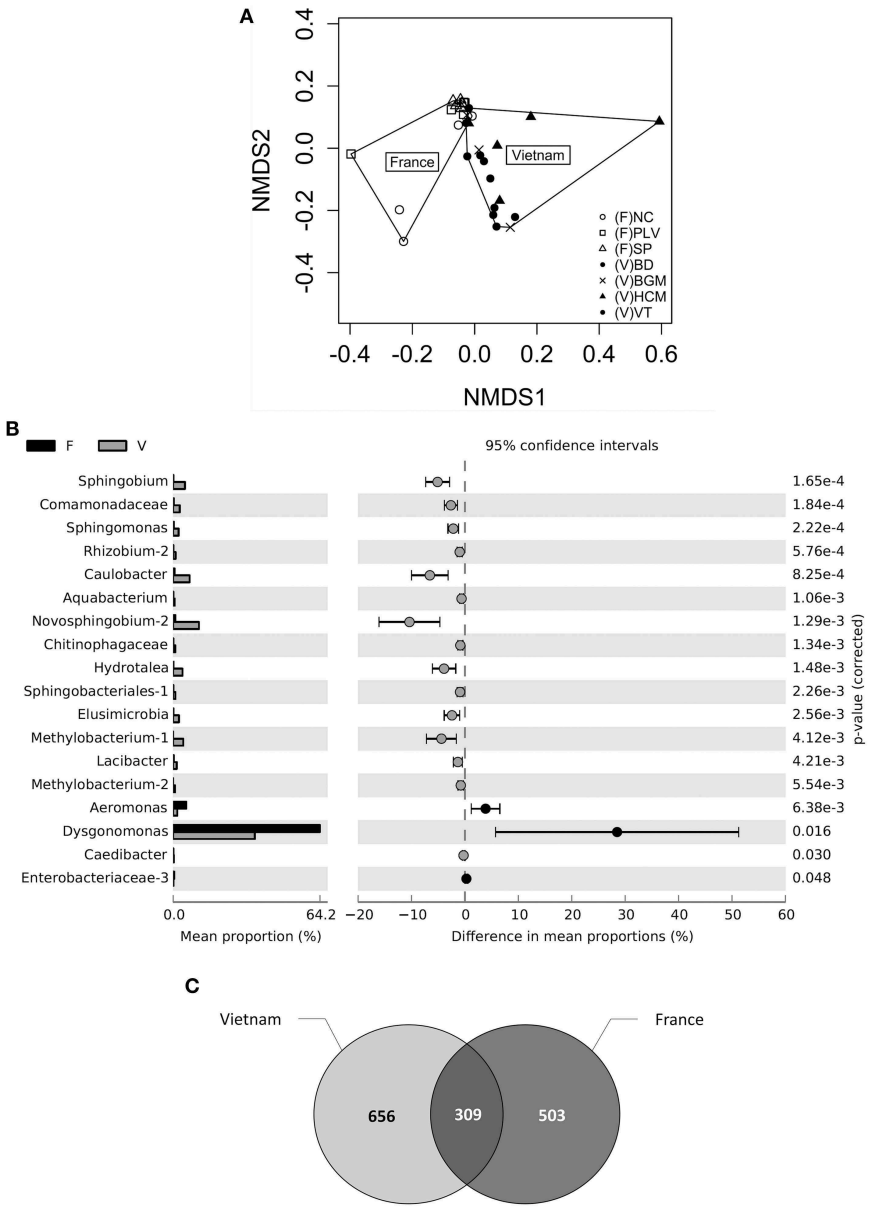



**FIGURE 4 | Community structure and OTU-country associations. (A)** Non-Metric Multidimentional Scaling plot represents Bray-Curtis $\beta$-diversity structure among individuals and populations. The loss function Stress = 0.09 and correlation with true distances $R^2 = 0.97$. NC, Nice; PLV, Porte-lès-Valence; SP, Saint Priest; VT, Vung Tàu City; HCM, Hồ Chí Minh City; BD, Bình Du'o'ng; BGM, Bù Gia Mâp. **(B)** Difference in OTU abundance between midgut bacterial microbiota of French and Vietnamese *Ae. albopictus*. Display represents extended error bars of significant fold-changes ($P > 0.05$) between midgut samples from France (F) and Vietnam (V). **(C)** Venn diagram representing shared OTUs between midgut samples from France and Vietnam. The intersection of both circles represents the number of shared OTUs between France and Vietnam. To avoid size effect, sequences were merged per group, and then subsampled.

ENA correction for null alleles and 0.145 ($CI_{95\%}$ 0.061–0.270) without correction. Consequently, the presence of null alleles did not strongly impact the estimation of differentiation. The structure of the populations was evaluated with the Bayesian method of assignment. The optimal number of clusters selected with the second-order change in likelihood method was $K = 2$ (**Figure S1**) (Evanno et al., 2005). Populations were clustered in two different genetic groups according to country of origin, except for populations from PLV in France and BD in Vietnam, which harbored a mixture of both genotypes (**Figure 5**, **Figure S5**). AMOVA analysis revealed a non-significant variation in the structure among countries (AMOVA 2.6%, $p = 0.2$) but a

**TABLE 2 | Microsatellite characteristics among populations.**

| Country | Site | Index | di-6 | tri-3 | tri-18 | tri-25 | tri-41 | tri-45 | tri-6 | B51 | B52 | A9 | AEDC | All |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Vietnam | BD | Nall | 0.03 | **0.21** | 0.03 | **0.16** | 0 | 0.07 | 0 | 0 | 0 | 0.01 | 0 | 0.046 |
| | | Fis | 0.114 | 0.49 | 0.125 | 0.366 | 0.058 | 0.211 | 0.038 | −0.036 | −0.14 | −0.008 | −0.164 | 0.133 |
| | | He | 0.812 | 0.828 | 0.92 | 0.798 | 0.834 | 0.746 | 0.852 | 0.95 | 0.259 | 0.813 | 0.593 | 0.717 |
| | | Ho | 0.733 | 0.433 | 0.82 | 0.519 | 0.8 | 0.6 | 0.833 | 1 | 0.3 | 0.833 | 0.7 | 0.636 |
| | HCM | Nall | 0 | **0.17** | 0.01 | **0.22** | 0.11 | 0.05 | 0.04 | 0.06 | 0 | 0.06 | 0.03 | 0.068 |
| | | Fis | −0.105 | 0.395 | 0.183 | 0.556 | 0.533 | 0.109 | 0.098 | 0.116 | na | 0.188 | 0.137 | 0.196 |
| | | He | 0.683 | 0.723 | 0.68 | 0.691 | 0.723 | 0.882 | 0.798 | 0.633 | 0 | 0.624 | 0.493 | 0.659 |
| | | Ho | 0.767 | 0.448 | 0.57 | 0.318 | 0.533 | 0.8 | 0.733 | 0.645 | 0 | 0.517 | 0.433 | 0.537 |
| | VT | Nall | 0 | 0.01 | **0.16** | **0.25** | **0.13** | 0 | 0 | 0 | 0 | 0.07 | 0 | 0.056 |
| | | Fis | −0.137 | 0.146 | 0.36 | 0.588 | 0.305 | 0.049 | −0.008 | na | −0.061 | 0.123 | −0.309 | 0.122 |
| | | He | 0.77 | 0.696 | 0.87 | 0.725 | 0.803 | 0.848 | 0.866 | 0 | 0.364 | 0.827 | 0.647 | 0.706 |
| | | Ho | 0.889 | 0.607 | 0.57 | 0.308 | 0.571 | 0.821 | 0.889 | 0 | 0.393 | 0.741 | 0.857 | 0.615 |
| | BGM | Nall | 0 | **0.19** | 0.01 | 0.02 | **0.08** | **0.13** | **0.11** | 0.05 | **0.12** | **0.16** | 0.01 | 0.08 |
| | | Fis | 0.026 | 0.429 | 0.012 | 0.071 | 0.218 | 0.267 | 0.264 | 0.119 | 0.267 | 0.369 | −0.309 | 0.171 |
| | | He | 0.775 | 0.693 | 0.9 | 0.7 | 0.791 | 0.783 | 0.84 | 0.386 | 0.542 | 0.838 | 0.65 | 0.729 |
| | | Ho | 0.773 | 0.409 | 0.91 | 0.667 | 0.636 | 0.591 | 0.636 | 0.35 | 0.41 | 0.546 | 0.86 | 0.627 |
| France | NC | Nall | 0.08 | 0.21 | 0.07 | 0.03 | 0 | 0.2 | **0.09** | 0 | 0 | 0.07 | 0 | 0.068 |
| | | Fis | 0.129 | 0.562 | 0.328 | 0.017 | −0.057 | 0.539 | 0.223 | na | −0.024 | 0.136 | −0.141 | 0.181 |
| | | He | 0.638 | 0.634 | 0.72 | 0.756 | 0.807 | 0.655 | 0.825 | 0 | 0.099 | 0.847 | 0.654 | 0.642 |
| | | Ho | 0.567 | 0.286 | 0.5 | 0.762 | 0.867 | 0.31 | 0.655 | 0 | 1.03 | 0.75 | 0.762 | 0.545 |
| | PLV | Nall | **0.08** | **0.13** | 0.06 | **0.12** | 0.01 | 0 | 0.05 | 0 | **0.11** | **0.22** | 0 | 0.071 |
| | | Fis | 0.207 | 0.507 | 0.108 | 0.211 | 0.033 | −0.143 | 0.13 | na | 0.487 | 0.51 | −0.306 | 0.152 |
| | | He | 0.494 | 0.292 | 0.77 | 0.727 | 0.745 | 0.316 | 0.827 | 0 | 0.127 | 0.686 | 0.504 | 0.551 |
| | | Ho | 0.4 | 0.148 | 0.7 | 0.607 | 0.733 | 0.367 | 0.733 | 0 | 0.067 | 0.345 | 0.667 | 0.475 |
| | SP | Nall | 0.04 | 0.03 | 0 | 0.09 | 0.05 | 0.01 | 0.08 | 0 | 0 | 0.05 | 0.06 | 0.037 |
| | | Fis | 0.005 | 0.154 | −0.012 | 0.223 | 0.122 | 0.011 | 0.191 | na | *na* | 0.2 | 0.176 | 0.12 |
| | | He | 0.579 | 0.4 | 0.84 | 0.765 | 0.757 | 0.48 | 0.71 | 0 | 0 | 0.675 | 0.574 | 0.706 |
| | | Ho | 0.586 | 0.345 | 0.86 | 0.607 | 0.679 | 0.483 | 0.586 | 0 | 0 | 0.552 | 0.483 | 0.615 |

*NC, Nice; PLV, Porte-lès-Valence; SP, Saint Priest; VT, Vung Tàu city; HCM, Hồ Chí Minh City; BD, Bình Du'ơng, BGM, Bù Gia Mập; Nall, null alleles frequency; Fis, fixation index; He, expected heterozygosity; Ho, observed heterozygosity. Significant frequencies of null alleles are in bold.*

moderate variation among sites within a country (AMOVA 13%, $p < 10^{-3}$) (**Table 1**).

### Positive Correlation between Bacterial and Genetic Diversities of *Ae. Albopictus* Populations

The populations sampled were compared to assess whether there was any relationship between the bacterial diversity and the genetic diversity of the mosquitoes. Comparative analysis showed a low correlation between bacterial β-diversity (Bray-Curtis dissimilarity distance) and haplotype structure (Mantel, $R^2 = 0.5$, $p = 0.02$) and no significant correlation with the Cavalli-Sforza Edwards measure of microsatellite genetic distance (Mantel, $R^2 = 0.20$, $p = 0.19$). However, for all sampling sites, positive correlations were observed between mean bacterial α-diversity ($H'$) and respectively host *Ar* (Spearman's rank correlation, $\rho = 0.95$, $p = 8.10^{-3}$) (**Figure 6A**) and host genetic diversity (*Hs*) (Spearman's rank correlation, $\rho = 0.78$, $p = 0.048$) (**Figure 6B**).

## Discussion

Mosquitoes may be regarded as holobiont units in which the host and its microbiota may display symbiotic relationships and multitrophic interactions (Minard et al., 2013). As demonstrated in other insect models such as drosophila and aphids (Engel and Moran, 2013; Lizé et al., 2014), it is envisaged that mosquito-associated bacterial microbiota could influence the ability of the host to respond to biotic and abiotic factors. An essential step forward is to increase our knowledge on the mosquito-associated microbiota. Here, we used a next generation sequencing method and metabarcoding to characterize the composition of bacterial microbiota of invasive and autochthonous populations of *Ae. albopictus* and to test for correlations with host genotype.

*Wolbachia* is the most studied bacterium of mosquitoes. Although, some beneficial fitness effects of this endosymbiont have been demonstrated in *Ae. albopictus* (Dobson et al., 2002), *Wolbachia* most commonly alters mosquito reproduction by

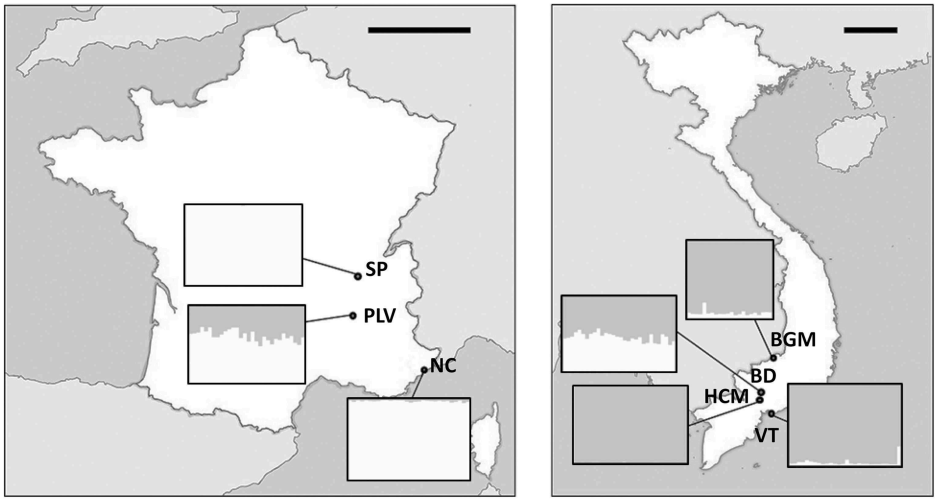



**FIGURE 5 | Genetic structures of *Aedes albopictus* populations.** The map of the microsatellite genetic structure ($K = 2$) for each site. Each bar represents an individual and the grayscale represents the probability that an individual belongs to a population. Scale bar of the maps, 200 km. NC, Nice; PLV, Porte-lès-Valence; SP, Saint Priest; VT, Vung Tàu City; HCM, Hồ Chí Minh City; BD, Bình Du'o'ng; BGM, Bù Gia Mập.

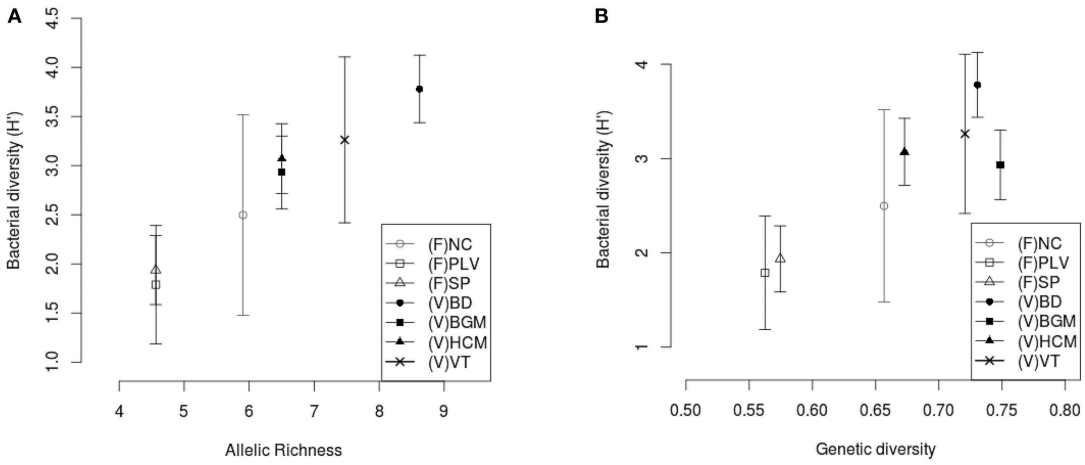



**FIGURE 6 | Correlation between host genetic richness, host genetic diversity, and midgut bacterial diversity.** The mean bacterial Shannon α-diversity (H′) was correlated with **(A)** rarefied genetic richness, Ar (Spearman's rank correlation, $\rho = 0.95$, $p = 8.10^{-3}$) and **(B)** diversity, Hs (Spearman's rank correlation, $\rho = 0.78$, $p = 0.048$). (F), France; (V), Vietnam; NC, Nice; PLV, Porte-lès-Valence; SP, Saint Priest; VT, Vung Tàu City; HCM, Hồ Chí Minh City; BD, Bình Du'o'ng; BGM, Bù Gia Mập. Standard deviation of Shannon α-diversity (H′) was represented for each site.

inducing cytoplasmic incompatibility between infected males and uninfected females (Stouthamer et al., 1999). Superinfection with more than one *Wolbachia* strain makes the system more complex and some models predict that superinfected females have a selective advantage as they should be able to reproduce with every combination of uninfected and strain-infected males (Dobson et al., 2004; Tortosa et al., 2010). Diagnostic PCR on field-caught *Ae. albopictus* from various countries confirmed this prediction as more than 99% of individuals were superinfected with *Wolbachia w*AlbA and *w*AlbB strains (Kittayapong et al., 2002; de Albuquerque et al., 2011; Zouache et al., 2011). Here we showed a 100% prevalence of such double infection in seven *Ae. albopictus* populations, three from France and four from Vietnam. Other than cytoplasmic incompatibility and the age of the mosquito (Tortosa et al., 2010), factors driving *Wolbachia* persistence in *Ae. albopictus* remain largely unknown. Our results demonstrated for the first time a strong correlation between the proportions of the two strains *w*AlbA and *w*AlbB in *Ae. albopictus* natural populations, suggesting a structured mechanism is regulating their co-occurrence. *Wolbachia* has also

been demonstrated to induce selective "sweep" on mitochondria, which drastically reduces the genomic variability of this organelle in *Ae. albopictus* (Hurst and Jiggins, 2005). Accordingly, the mitochondrial haplotypes of all seven populations showed very low nucleotide diversity in the *COI* gene (haplotypes differed by only one substitution) and low richness (less than three haplotypes per sampling site after rarefaction). Therefore, mitochondrial markers are highly sensitive to admixture in comparison with nuclear markers such as microsatellites. For this reason, it is not suitable to rely on mitochondrial haplotypes when estimating intraspecific genetic diversity in *Ae. albopictus*.

As confirmed herein, analyses of the *Ae. albopictus* microbiota are systematically dominated by *Wolbachia* sequences (Minard et al., 2014). In order to investigate other genera we used the midgut tissue, which is recognized to be poorly colonized by this bacterium (Zouache et al., 2009). Moreover, in arthropod vectors, this organ is the main site for multipartite interactions between bacterial microbiota, arboviruses and the host (Jupatanakul et al., 2014; Kenney and Brault, 2014a). Interestingly, 21% of total OTUs richness was found in populations from both France and Vietnam, assuming that some members of the microbiota can be shared among *Ae. albopictus* from contrasted populations. Despite the relatively low richness of shared OTUs, they accounted for 85.2% of all the sequences. The dominance of the shared microbiota over the transient microbiota is suggestive of positive selection of the interactions that constitute semi-constant gut microbiota (**Figure S3**), in contrast with the variable microbiota previously described within midguts of field-caught mosquitoes belonging to different species (Osei-Poku et al., 2012). Although, this observation may also be explained by the relatively homogeneity of the habitats of *Aedes* spp. mosquitoes (low oxygen pressure, a pH comprised between 8 and 10) (Dillon and Dillon, 2004; del Pilar Corena et al., 2005; Saboia-Vahia et al., 2014), and the establishment of habitat-specific bacterial associations. However, some factors such as the type of nutrients ingested and temperature (not regulated in ectotherms) can strongly vary between French and Vietnamese populations. In addition, previous studies highlighted the importance of water of breeding sites in which mosquito larvae and pupae develop. It was shown that mosquitoes acquire a large part of their microbiota from larval stages which themselves depend on the bacterial composition of the water of breeding sites. Moreover, a great variability in diversity of abundance of taxa was shown between stages according to mosquito species (Pumpuni et al., 1996; Minard et al., 2013; Coon et al., 2014; Dada et al., 2014; Gimonneau et al., 2014). Following these observations, the variations observed among the populations studied here could be linked with environmental factors of their habitat.

Among all the midgut samples, a total of 68 dominant OTUs were described and assigned with the name of their most probable phylotype. *Dysgonomonas* was the most prevalent and abundant one. This genus belongs to the phylum *Bacteroidetes*, and has already been recently detected in *Ae. albopictus* from Madagascar (Minard et al., 2014). Prevalent and abundant bacteria belonging to the *Bacteroidetes* phylum were previously described in *Ae. aegypti*. In particular, this mosquito species harbors a high relative abundance of *Chryseobacterium* that is maintained during all different life stages of lab-reared populations (Coon et al., 2014). However, previous studies highlighted that gut mosquito microbiota is largely dominated by *Proteobacteria* as for *Anopheles coluzzi*, *An. funestus*, *An. gambiae*, *An. Stephensi,* and *Culex tarsalis* (Pidiyar et al., 2004; Lindh et al., 2005; Rani et al., 2009; Boissière et al., 2012; Minard et al., 2013; Gimonneau et al., 2014). For these studies one to two major genera were usually found dominant in the midgut, albeit the identity of a given dominant entity changes over individuals or populations (Boissière et al., 2012; Osei-Poku et al., 2012). Therefore, it is surprising that we describe a single prevalent and abundant taxon, *Dysgonomonas*, in the midgut of all individuals even originated from distantly separated populations, suggesting an evolutionary process that maintains the presence of this particular taxon. Interestingly, *Dysgonomonas* has also been identified in the gut of different animals such as termites (*Coptotermes formosanus, Macrotermes barneyi*), house flies (*Musca domestica*), fruit flies (*Drosophila* spp.), red palm weevils (*Rhynchophorus ferrugineus*), and sea bass (*Dicentrarchus labrax*) (Husseneder et al., 2009; Chandler et al., 2011; Carda-Diéguez et al., 2014; Tagliavia et al., 2014; Yang et al., 2014). This genus has been previously detected in *Anopheles stephensi* and *Culex tarsalis* microbiota with a moderate abundance (Rani et al., 2009; Duguma et al., 2013). Its ability to cause lysis of erythrocytes and to synthesize $B_{12}$ vitamins might be a selective mechanism involved in a mutualistic interaction with female mosquitoes (Hironaga et al., 2008; Husseneder et al., 2009; Lawson et al., 2010; Yang et al., 2014). Some species of *Dysgonomonas* are both aerobic and facultative anaerobic, which could partly explain how some may adapt to the nearly anoxic insect midgut habitat (Johnson and Barbehenn, 2000; Chouaia et al., 2014). All these observations point to *Dysgonomonas* having a role in mosquito biology. Finally, certain *Dysgonomonas* species were also identified as human opportunistic pathogens, raising concern about the possibility of additional mosquito borne diseases and thus highlighting the recent concept of "pathobiome" in arthropod vectors (Hironaga et al., 2008; Lawson et al., 2010; Vayssier-Taussat et al., 2014).

Other bacterial genera detected in *Ae. albopictus* belong to the *Sphingomonadaceae* family (*Sphingomonas*, *Sphingobium*, *Novosphingobium*) and were preferentially associated with the autochthonous populations from Vietnam. These bacterial genera are ubiquitous in the environment (Vaz-Moreira et al., 2011; Ashton Acton, 2012) and are also able to colonize a variety of higher organisms (D'Auria et al., 2013; Zhang et al., 2013; Dai et al., 2014). They display various catabolic abilities including the production of hydrolases involved in the degradation of oligosaccharides, and in termite hosts they may participate in the degradation of plant compounds (Aylward et al., 2013). Moreover, these bacteria have already been identified in other plant-feeding insects including mosquitoes *Anopheles maculipenis*, *Anopheles gambiae*, *Anopheles stephensis,* and *Aedes aegypti* (Dong et al., 2009; Ramírez-Puebla et al., 2010; Dinparast Djadid et al., 2011; Gayatri Priya et al., 2012; Terenius et al., 2012; Koroiva et al., 2013). From these observations, it could be assumed that *Sphingomonadaceae* are important in making plant sugar available to the mosquito host, by degrading

oligosaccharides in the mosquito gut or acquiring them from the environment.

Interestingly, the bacterial communities associated with the invasive mosquito populations in France were less diverse and more homogeneous than those associated with autochthonous populations in Vietnam. The living host environment is an important factor impacting the microbiota of insects (Linnenbrink et al., 2013). In mosquitoes, it is known that feeding behavior may drastically modify the structure of gut microbiota and strongly increase inter-individual variation (Wang et al., 2011; Pernice et al., 2014). To avoid the effects of short-term changes in midgut microbiota, we only studied unfed mosquitoes, but long-term feeding effects cannot be ignored. Indeed, mosquito nutrition is mostly based on nectar and *Ae. albopictus* is known to have a wide nutritional spectrum (Clements, 1992). Interestingly, previous studies of the microflora diversity of different environments highlighted a significant reduction of species richness within the flowering plants (angiosperms) in temperate compared to tropical ecosystems (Francis and Currie, 2003). Therefore, the diversity and availability of plant nutrient sources could explain the reduction and the homogeneity in bacterial diversity we observed in mosquito populations of France. Populations from France also harbor a lower genetic diversity. Genetic reductions in invasive species have been widely documented. The most probable factor explaining genetic reduction in invasive populations is that recent colonization by a reduced effective population size causes a founder effect bottleneck and genetic drift (Dlugosch and Parker, 2008). However, no evidence for a recent founder effect was detected in the invasive populations in France. The short generation time of *Ae. albopictus* as well as the history of complex and multiple introductions (evidenced especially in Portes-Lès-Valence populations which harbor low allelic richness and genetic diversity but an admixture structure discovered with both haplotype and microsatellite markers) may have erased signs of a past bottleneck. In addition, genetic diversity reductions in France could also be explained by other patterns (e.g. Landscape fragmentation, lowest effective size…). Interestingly, the most genetic diversified of these populations was in Nice, a invaded site since 2004 (Medlock et al., 2012), whereas Portes-Lès-Valence and Saint-Priest were colonized much later. As already suggested in various models (Sommer, 2005), genetic diversity reduction identified with neutral markers could be linked with diversity reduction of genes involved in immunity. Moreover, immune genes could also be under considerable selective pressure that would affect the composition of mosquito microbiota (Wang et al., 2011; Minard et al., 2013). Indeed, gut microbiota are involved in a strong reciprocal interaction with the host immune system in the mosquito gut (Hillyer, 2010; Cirimotich et al., 2011; Weiss and Aksoy, 2011). However, we did not find any correlations between microbiota and mosquito genetic structure based on neutral microsatellite markers. Interestingly, a genetic study of mice using quantitative markers demonstrated that a core microbiome was regulated by a complex polygenic trait likely to involve pleiotropic effects (Benson et al., 2010). However, as a direct link between genetic and microbial structures cannot be proven following our sampling design, further investigations would be necessary. In particular, development of *Ae. albopictus* quantitative markers would be helpful in pinpointing which host genetic factors could partly shape the microbiota diversity.

Reduction in diversity for both the mosquito host and its associated bacterial microbiota also raises questions about the possible impact on human pathogen transmission. During transmission cycles, mosquito-vectored pathogens pass through the host midgut epithelial membrane to reach hemolymph and salivary glands. This checkpoint is critical for transmission as the pathogen faces the microbiota barrier and its potentially antagonistic activity (enzymes, toxins, etc.) as well as the host immune system (Cirimotich et al., 2011; Weiss and Aksoy, 2011; Wang et al., 2013). Indeed, the microbiota of mosquito vectors was shown to interfere, positively or negatively, with their susceptibility to pathogen infection and transmission capacity (Dennison et al., 2014). For instance, a recent study on *Anopheles gambiae* mosquitoes suggested that reduction of gut microbiota diversity following ingestion of antibiotics increases the capacity of females to transmit *Plasmodium falciparum* (Gendrin et al., 2015). In previous works it was demonstrated that a high proportion of viral particles were able to disseminate beyond the insect midgut barrier in *Ae. albopictus* populations from Cagne-Sur-Mer (∼12 km away from Nice in France) (Vega-Rua et al., 2013) whereas the lowest range of Chikungunya virus dissemination was found in mosquitoes from Vietnam (Zouache et al., 2014). Indeed, the midgut microbiota is the first barrier encountered by viruses during their infection process and its diversity could strongly interferes with virus replication (Jupatanakul et al., 2014; Kenney and Brault, 2014a). However, microbiota is not the only factor that can modulate mosquito competence. Vector spatial genetic structure (gene flow, presence of cryptic species, invasion) has previously been demonstrated to greatly influence its interactions with pathogens (reviewed by McCoy, 2008; Léger et al., 2013). In *Ae. albopictus* recent studies highlighted that complex genetic-genetic-environment interactions impacted the transmission of Chikungunya virus (Zouache et al., 2014). However, the effect of mosquito genetic diversity on disease transmission remains unclear. Nevertheless, previous empirical evidences and models based on host-pathogens systems predicted that host genetic diversity can negatively affect the prevalence of pathogens (reviewed by King and Lively, 2012). It is conceivable that the reduction and change in the mosquito microbiota and its associated immune response could partly explain the efficient vector competence observed in *Ae. albopictus* populations from Metropolitan France.

In conclusion, our results suggest a similar pattern in reduction of genetic diversity of *Ae. albopictus* and bacterial microbiota diversity. This finding provides new insights into the biology of an invasive species and its associated bacterial microbiota. It also highlights the need for further ecological studies to describe how the invasive mosquito population, as well as its hologenome, responds when challenged by new biotic and abiotic factors. Moreover, the dynamics of mosquito-associated eukaryotic and viral microbiota should also be investigated to gain a fully integrated view of the holobiont pathosystem of the Asian tiger mosquito.

## Ethical Issues

No ethical issues to be promulgated.

## Author Contributions

This work is part of GM's PhD dissertation (supervised by PM and CV) on *Aedes albopictus* microbiota. GM, PM, and CV conceived the project and sampling design. CB, TH, GL, KL, GM, PM, CV, and VT contributed to collection of specimens. GM, CV, FT, and VT performed all molecular work and genotyping scoring. GM and CG analyzed and interpreted the genotypic data and GM analyzed and interpreted the metabarcoding data. GM wrote the article and all other authors contributed edits and comments.

## Funding

Funding for this project was provided by grants from EC2CO CNRS and CMIRA Région Rhône-Alpes. This research was also partially funded by ERA-NET BiodivERsA with the national funders ANR-13-EBID-0007-01, FWF I-1437, and DFG KL 2087/6-1 as part of the 2012-2013 BiodivERsA call for research proposals, and was carried out within the framework of GDRI "Biodiversity and Infectious Diseases in Southeast Asia." To achieve this work, we used the computing facilities of the PRABI cluster. We also gratefully acknowledge the contribution of the DTAMB platform of the FR41 Bio-Environment and Health (University Lyon 1).

## Acknowledgments

We thank help by Bernard Kaufmann with population genetics advices, Patricia Luis with map design, Audrey Dubost with bioinformatics analysis, Guillaume Carillo with qPCR experiments, and David Wilkinson for reading the revised version of the manuscript. We also thank the two anonymous reviewers for their helpful comments on the first version of the manuscript.

## Supplementary Material

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fmicb.2015.00970

**Figure S1 | Prediction of the best value of *K*. According to Evanno et al. (2005), the distribution of Δ*K* (absolute values of the second-order in change of the likelihood distribution divided by the standard deviation of the likelihoods) was plotted for each value of K (number of potential sub-groups) from 2 to 7.**

**Figure S2 | Index comparisons between French and Vietnamese populations according to the 3% OTU richness estimations (Chao1), α-diversity (H') and β-diversity homogeneity (Bray-Curtis Distance to centroid).**

**Figure S3 | Prevalence of Operational Taxonomic Units (OTUs) according to their mean relative abundance in midgut samples.** Most abundant OTUs (proportion > 0.01) are named by their assignation according to naïve Bayesian classifier (Bootstrap > 80%). The prevalence and abundance of OTUs were calculated from sequences obtained from midgut samples analyzed in the study.

**Figure S4 | Map of haplotypes.** The six haplotypes proportion (H_1, H_2, H_3, H_4, H_5, H_6) are represented for each site. Scale bar, 200 km. NC, Nice; PLV, Porte-lès-Valence; SP, Saint Priest; VT, Vung Tàu City; HCM, Hồ Chí Minh City; BD, Bình Du'o'ng; BGM, Bù Gia Mâp.

**Figure S5 | Factorial Correspondence Analysis of mosquitos' genetic structure.** Each point represents an individual from a given population. The color indicates different original population of individuals. The ordination is based on multivariate analysis of allelic frequencies of the 11 microsatellites among individuals. NC, Nice; PLV, Porte-lès-Valence; SP, Saint Priest; VT, Vung Tàu City; HCM, Hồ Chí Minh City; BD, Bình Du'o'ng; BGM, Bù Gia Mâp.

**Table S1 | Sample compositions for the different analyses.**

**Table S2 | Microsatellite primers and information.**

**Table S3 | AMOVA analysis of phylogeny based Unifrac $\beta$-diversity.**

**Table S4 | Haplotypes and nucleotide diversity.**

**Table S5 | Bottleneck analysis.**

**Table S6 | Dominant contaminant OTUs found in the negative control.**

## References

Andersson, A. F., Lindberg, M., Jakobsson, H., Bäckhed, F., Nyrén, P., and Engstrand, L. (2008). Comparative analysis of human gut microbiota by barcoded pyrosequencing. *PLoS ONE* 3:e2836. doi: 10.1371/journal.pone.0002836

Ashton Acton, Q. (2012). *Advances in Sphingomonadaceae Research and Application.* Atlanta, GA: ScholarlyEditions.

Aylward, F. O., McDonald, B. R., Adams, S. M., et al. (2013). Comparison of 26 Sphingomonad genomes reveals diverse environmental adaptations and biodegradative capabilities. *Appl. Environ. Microbiol.* 79, 3724–3733. doi: 10.1128/AEM.00518-13

Beebe, N. W., Ambrose, L., Hill, L. A., Davis, J. B., Hapgood, G., Cooper, R. D., et al. (2013). Tracing the tiger: population genetics provides valuable insights into the *Aedes* (Stegomyia) *albopictus* invasion of the Australasian region. *PLoS Negl. Trop. Dis.* 7:e2361. doi: 10.1371/journal.pntd.0002361

Benson, A. K., Kelly, S. A., Legge, R., Ma, F., Low, S. J., Kim, J., et al. (2010). Individuality in gut microbiota composition is a complex polygenic trait shaped by multiple environmental and host genetic factors. *Proc. Natl. Acad. Sci. U.S.A.* 107, 18933–18938. doi: 10.1073/pnas.1007028107

Blottière, H. M., de Vos, W. M., Ehrlich, S. D., and Doré, J. (2013). Human intestinal metagenomics: state of the art and future. *Curr. Opin. Microbiol.* 16, 232–239. doi: 10.1016/j.mib.2013.06.006

Boissière, A., Tchioffo, M. T., Bachar, D., Abate, L., Marie, A., Nsango, S. E., et al. (2012). Midgut microbiota of the malaria mosquito vector *Anopheles gambiae* and interactions with *Plasmodium falciparum* infection. *PLoS Pathog.* 8:e1002742. doi: 10.1371/journal.ppat.1002742

Bonizzoni, M., Gasperi, G., Chen, X., and James, A. A. (2013). The invasive mosquito species *Aedes albopictus*: current knowledge and future perspectives. *Trends Parasitol.* 29, 460–468. doi: 10.1016/j.pt.2013.07.003

Buchner, P. (1965). *Endosymbiosis of Animals with Plant Microorganisms*. Bucks, PA: Interscience Publishers.

Carda-Diéguez, M., Mira, A., and Fouz, B. (2014). Pyrosequencing survey of intestinal microbiota diversity in cultured sea bass (*Dicentrarchus labrax*) fed functional diets. *FEMS Microbiol. Ecol.* 87, 451–459. doi: 10.1111/1574-6941.12236

Chambers, D. M., Young, L. F., and Hill, H. S. (1986). Backyard mosquito larval habitat availability and use as influenced by census tract determined resident income levels. *J. Am. Mosq. Control Assoc.* 2, 539–544.

Chandler, J. A., Lang, J. M., Bhatnagar, S., Eisen, J. A., and Kopp, A. (2011). Bacterial communities of diverse *Drosophila* species: ecological context of a host-microbe model system. *PLoS Genet.* 7:e1002272. doi: 10.1371/journal.pgen.1002272

Chapuis, M.-P., and Estoup, A. (2007). Microsatellite null alleles and estimation of population differentiation. *Mol. Biol. Evol.* 24, 621–631. doi: 10.1093/molbev/msl191

Chouaia, B., Gaiarsa, S., Crotti, E., Comandatore, F., Degli Esposti, M., Ricci, I., et al. (2014). Acetic acid bacteria genomes reveal functional traits for adaptation to life in insect guts. *Genome Biol. Evol.* 6, 912–920. doi: 10.1093/gbe/evu062

Cirimotich, C. M., Ramirez, J. L., and Dimopoulos, G. (2011). Native microbiota shape insect vector competence for human pathogens. *Cell Host Microbe* 10, 307–310. doi: 10.1016/j.chom.2011.09.006

Clements, A. N. (1992). *The Biology of Mosquitoes: Development, Nutrition and Reproduction.* London: Chapman & Hall.

Coon, L. K., Vogel, J. K., Brown, M. R., and Strand, M. R. (2014). Mosquitoes rely on their gut microbiota for development. *Mol. Ecol.* 23, 2727–2739. doi: 10.1111/mec.12771

Cornuet, J. M., and Luikart, G. (1996). Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data. *Genetics* 144, 2001–2014.

D'Auria, G., Peris-Bondia, F., DŽunková, M., Mira, A., Collado, M. C., Latorre, A., et al. (2013). Active and secreted IgA-coated bacterial fractions from the human gut reveal an under-represented microbiota core. *Sci. Rep.* 3:3515. doi: 10.1038/srep03515

Dada, N., Jumas-Bilak, E., Manguin, S., Seidu, R., Stenström, T. A., and Overgaard, H. J. (2014). Comparative assessment of the bacterial communities associated with *Aedes aegypti* larvae and water from domestic water storage containers. *Parasit. Vectors* 7:391. doi: 10.1186/1756-3305-7-391

Dai, J. X., Liu, X. M., and Wang, Y. J. (2014). Diversity of endophytic bacteria in *Caragana microphylla* grown in the desert grassland of the Ningxia Hui autonomous region of China. *Genet. Mol. Res.* 13, 2349–2358. doi: 10.4238/2014.April.3.7

de Albuquerque, A. L., Magalhães, T., and Ayres, C. F. J. (2011). High prevalence and lack of diversity of *Wolbachia* pipientis in *Aedes albopictus* populations from Northeast Brazil. *Memórias do Instituto Oswaldo Cruz* 106, 773–776. doi: 10.1590/S0074-02762011000600021

del Pilar Corena, M., VanEkeris, L., Salazar, M. I., Bowers, D., Fiedler, M. M., Silverman, D., et al. (2005). Carbonic anhydrase in the adult mosquito midgut. *J. Exp. Biol.* 208, 3263–3273. doi: 10.1242/jeb.01739

Dennison, N. J., Jupatanakul, N., and Dimopoulos, G. (2014). The mosquito microbiota influences vector competence for human pathogens. *Curr. Opin. Insect Sci.* 3, 6–13. doi: 10.1016/j.cois.2014.07.004

Dillon, R. J., and Dillon, V. M. (2004). The gut bacteria of insects: nonpathogenic interactions. *Annu. Rev. Entomol.* 49, 71–92. doi: 10.1146/annurev.ento.49.061802.123416

Dinparast Djadid, N., Jazayeri, H., Raz, A., Favia, G., Ricci, I., and Zakeri, S. (2011). Identification of the midgut microbiota of *An. stephensi* and *An. maculipennis* for their application as a paratransgenic tool against malaria. *PloS ONE* 6:e28484. doi: 10.1371/journal.pone.0028484

Dlugosch, K. M., and Parker, I. M. (2008). Founding events in species invasions: genetic variation, adaptive evolution, and the role of multiple introductions. *Mol. Ecol.* 17, 431–449. doi: 10.1111/j.1365-294X.2007.03538.x

Dobson, S. L., Marsland, E. J., and Rattanadechakul, W. (2002). Mutualistic *Wolbachia* infection in *Aedes albopictus*: accelerating cytoplasmic drive. *Genetics* 160, 1087–1094.

Dobson, S. L., Rattanadechakul, W., and Marsland, E. J. (2004). Fitness advantage and cytoplasmic incompatibility in *Wolbachia* single- and superinfected *Aedes albopictus*. *Heredity* 93, 135–142. doi: 10.1038/sj.hdy.6800458

Dong, Y., Manfredini, F., and Dimopoulos, G. (2009). Implication of the mosquito midgut microbiota in the defense against malaria parasites. *PLoS Pathog.* 5:e1000423. doi: 10.1371/journal.ppat.1000423

Douglas, A. E. (2011). Lessons from studying insect symbioses. *Cell Host Microbe* 10, 359–367. doi: 10.1016/j.chom.2011.09.001

Douglas, A. E. (2014). The molecular basis of bacterial-insect symbiosis. *J. Mol. Biol.* 426, 3830–3837. doi: 10.1016/j.jmb.2014.04.005

Dray, S., and Dufour, A. B. (2007). The ade4 Package: implementing the duality diagram for ecologists. *J. Stat. Softw.* 22, 1–20.

Duguma, D., Rugman-Jones, P., Kaufman, M. G., Hall, M. W., Neufeldd, J. D., Southamer, R., et al. (2013). Bacterial communities associated with *Culex* mosquito larvae and two emergent aquatic plants of bioremediation importance. *PLoS ONE* 8:e72522. doi: 10.1371/journal.pone.0072522

Earl, D. A., and VonHoldt, B. M. (2012). STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conserv. Genet. Resour.* 4, 359–361. doi: 10.1007/s12686-011-9548-7

Engel, P., and Moran, N. A. (2013). The gut microbiota of insects – diversity in structure and function. *FEMS Microbiol. Rev.* 37, 699–735. doi: 10.1111/1574-6976.12025

Evanno, G., Regnaut, S., and Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol. Ecol.* 14, 2611–2620. doi: 10.1111/j.1365-294X.2005.02553.x

Excoffier, L., and Lischer, H. E. L. (2010). Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Resour.* 10, 564–567. doi: 10.1111/j.1755-0998.2010.02847.x

Francis, A. P., and Currie, D. J. (2003). A globally consistent richness-climate relationship for angiosperms. *Am. Nat.* 161, 523–536. doi: 10.1086/368223

Gayatri Priya, N., Ojha, A., Kajla, M. K., Raj, A., and Rajagopal, R. (2012). Host plant induced variation in gut bacteria of *Helicoverpa armigera*. *PLoS ONE* 7:e30768. doi: 10.1371/journal.pone.0030768

Gendrin, M., Rodgers, F. H., Yerbanga, R. S., Ouédraogo, J. B., Basáñez, M. G., Cohuet, A., et al. (2015). Antibiotics in ingested human blood affect the mosquito microbiota and capacity to transmit malaria. *Nat. Commun.* 6:5921. doi: 10.1038/ncomms6921

Gimonneau, G., Tchioffo, M. T., Abate, L., Boissière, A., Awono-Ambéné, P. H., Nsango, S. E., et al. (2014). Composition of Anopheles coluzzii and Anopheles gambiae microbiota from larval to adult stages. *Infect. Genet. Evol.* 28, 715–724. doi: 10.1016/j.meegid.2014.09.029

Handley, L.-J. L., Estoup, A., Evans, D. M., Thomas, C. E., Lombaert, E., Facon, B., et al. (2011). Ecological genetics of invasive alien species. *Biocontrol* 56, 409–428. doi: 10.1007/s10526-011-9386-2

Hillyer, J. F. (2010). Mosquito immunity. *Adv. Exp. Med. Biol.* 708, 218–238. doi: 10.1007/978-1-4419-8059-5_12

Hironaga, M., Yamane, K., Inaba, M., Haga, Y., and Arakawa, Y. (2008). Characterization and antimicrobial susceptibility of *Dysgonomonas capnocytophagoides* isolated from human blood sample. *Jpn. J. Infect. Dis.* 61, 212–213.

Hurst, G. D. D., and Jiggins, F. M. (2005). Problems with mitochondrial DNA as a marker in population, phylogeographic and phylogenetic studies: the effects of inherited symbionts. *Proc. R. Soc. Lond. B Biol. Sci.* 272, 1525–1534. doi: 10.1098/rspb.2005.3056

Husseneder, C., Berestecky, J. M., and Grace, J. K. (2009). Changes in composition of culturable bacteria community in the gut of the Formosan subterranean termite depending on rearing conditions of the host. *Ann. Entomol. Soc. Am.* 102, 498–507. doi: 10.1603/008.102.0321

Jakobsson, M., and Rosenberg, N. A. (2007). CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* 23, 1801–1806. doi: 10.1093/bioinformatics/btm233

Johnson, K. S., and V Barbehenn, R. (2000). Oxygen levels in the gut lumens of herbivorous insects. *J. Insect Physiol.* 46, 897–903. doi: 10.1016/S0022-1910(99)00196-1

Jupatanakul, N., Sim, S., and Dimopoulos, G. (2014). The insect microbiome modulates vector competence for arboviruses. *Viruses* 6, 4294–4313. doi: 10.3390/v6114294

Kenney, J. L., and Brault, A. C. (2014a). *Advances in Virus Research.* Amsterdam: Academic Press.

Kenney, J. L., and Brault, A. C. (2014b). The role of environmental, virological and vector interactions in dictating biological transmission of arthropod-borne viruses by mosquitoes. *Adv. Virus Res.* 89, 39–83. doi: 10.1016/B978-0-12-800172-1.00002-1

King, K. C., and Lively, C. M. (2012). Does genetic diversity limit disease spread in natural host populations? *Heredity* 109, 199–203. doi: 10.1038/hdy.2012.33

Kittayapong, P., Baisley, K. J., Sharpe, R. G., Baimai, V., and O'Neill, S. L. (2002). Maternal transmission efficiency of *Wolbachia* superinfections in *Aedes albopictus* populations in Thailand. *Am. J. Trop. Med. Hyg.* 66, 103–107.

Koroiva, R., Souza, C. W. O., Toyama, D., Henrique-Silva, F., and Fonseca-Gessner, A. A. (2013). Lignocellulolytic enzymes and bacteria associated with the digestive tracts of *Stenochironomus* (Diptera: Chironomidae) larvae. *Genetic. Mol. Res.* 12, 3421–3434. doi: 10.4238/2013.April.2.2

Lawson, P. A., Carlson, P., Wernersson, S., Moore, E. R. B., and Falsen, E. (2010). *Dysgonomonas hofstadii* sp. nov., isolated from a human clinical source. *Anaerobe* 16, 161–164. doi: 10.1016/j.anaerobe.2009.06.005

Léger, E., Vourc'h, G., Vial, L., Chevillon, C., and McCoy, D. K. (2013). Changing distributions of ticks: causes and consequences. *Exp. Appl. Acarol.* 59, 9615. doi: 10.1007/s10493-012-9615-0

Librado, P., and Rozas, J. (2009). DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 25, 1451–1452. doi: 10.1093/bioinformatics/btp187

Lindh, J. M., Terenius, O., and Faye, I. (2005). 16S rRNA gene-based identification of midgut bacteria from field-caught *Anopheles gambiae* sensu lato and *Anopheles funestus* mosquitoes reveals new species related to known insect symbionts. *Appl. Environ. Microbiol.* 71, 7217–7223. doi: 10.1128/AEM.71.11.7217-7223.2005

Linnenbrink, M., Wang, J., Hardouin, E. A., Künzel, S., Metzler, D., and Baines, J. F. (2013). The role of biogeography in shaping diversity of the intestinal microbiota in house mice. *Mol. Ecol.* 22, 1904–1916. doi: 10.1111/mec.12206

Lizé, A., McKay, R., and Lewis, Z. (2014). Kin recognition in *Drosophila*: the importance of ecology and gut microbiota. *ISME J.* 8, 469–477. doi: 10.1038/ismej.2013.157

Masella, A. P., Bartram, A. K., Truszkowski, J. M., Brown, D. G., and Neufeld, J. D. (2012). PANDAseq: paired-end assembler for Illumina sequences. *BMC Bioinformatics* 13:31. doi: 10.1186/1471-2105-13-31

McCoy. (2008). The population genetic structure of vectors and our understanding of disease epidemiology. *Parasite* 15, 444–448. doi: 10.1051/parasite/2008153444

Medlock, J. M., Hansford, K. M., Schaffner, F., Versteirt, V., Hendrickx, G., and Zeller, H., et al. (2012). A review of the invasive mosquitoes in Europe: ecology, public health risks, and control options. *Vector Borne Zoonotic Dis.* 12, 435–447. doi: 10.1089/vbz.2011.0814

Meusnier, I., Olsen, J. L., Stam, W. T., Destombe, C., and Valero, M. (2001). Phylogenetic analyses of *Caulerpa taxifolia* (Chlorophyta) and of its associated bacterial microflora provide clues to the origin of the Mediterranean introduction. *Mol. Ecol.* 10, 931–946. doi: 10.1046/j.1365-294X.2001.01245.x

Minard, G., Mavingui, P., and Moro, C. V. (2013). Diversity and function of bacterial microbiota in the mosquito holobiont. *Parasit. Vectors* 6:146. doi: 10.1186/1756-3305-6-146

Minard, G., Tran, F.-H., Dubost, A., Tran-Van, V., Mavingui, P., and Moro, C. V. (2014). Pyrosequencing 16S rRNA genes of bacteria associated with wild tiger mosquito *Aedes albopictus*: a pilot study. *Front. Cell. Infect. Microbiol.* 4:59. doi: 10.3389/fcimb.2014.00059

Moran, N. A., McCutcheon, J. P., and Nakabachi, A. (2008). Genomics and evolution of heritable bacterial symbionts. *Annu. Rev. Genet.* 42, 165–190. doi: 10.1146/annurev.genet.41.110306.130119

Moya, A., Peretó, J., Gil, R., and Latorre, A. (2008). Learning how to live together: genomic insights into prokaryote-animal symbioses. *Nat. Rev. Genet.* 9, 218–229. doi: 10.1038/nrg2319

Muegge, B. D., Kuczynski, J., Knights, D., Clemente, J. C., González, A., and Fontana, L., et al. (2011). Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans. *Science* 332, 970–974. doi: 10.1126/science.1198719

Ochman, H., Worobey, M., Kuo, C.-H., Ndjango, J. B. N., Peeters, M., Hahn, B. H., et al. (2010). Evolutionary relationships of wild hominids recapitulated by gut microbial communities. *PLoS Biol.* 8:e1000546. doi: 10.1371/journal.pbio.1000546

Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'Hara, R. B., et al. (2013). *vegan: Community Ecology Package.* Available online at: https://cran.r-project.org/web/packages/vegan/index.html

Osei-Poku, J., Mbogo, C. M., Palmer, W. J., and Jiggins, F. M. (2012). Deep sequencing reveals extensive variation in the gut microbiota of wild mosquitoes from Kenya. *Mol. Ecol.* 21, 5138–5150. doi: 10.1111/j.1365-294X.2012.05759.x

Parks, D. H., and Beiko, R. G. (2010). Identifying biologically relevant differences between metagenomic communities. *Bioinformatics* 26, 715–721. doi: 10.1093/bioinformatics/btq041

Paupy, C., Delatte, H., Bagny, L., Corbel, V., and Fontenille, D. (2009). *Aedes albopictus*, an arbovirus vector: from the darkness to the light. *Microbes Infect.* 11, 1177–1185. doi: 10.1016/j.micinf.2009.05.005

Pernice, M., Simpson, S. J., and Ponton, F. (2014). Towards an integrated understanding of gut microbiota using insects as model systems. *J. Insect Physiol.* 69, 12–20. doi: 10.1016/j.jinsphys.2014.05.016

Pidiyar, V. J., Jangid, K., Patole, M. S., and Shouche, Y. S. (2004). Studies on cultured and uncultured microbiota of wild *Culex quinquefasciatus* mosquito midgut based on 16s ribosomal RNA gene analysis. *Am. J. Trop. Med. Hyg.* 70, 597–603.

Porretta, D., Gargani, M., Bellini, R., Calvitti, M., and Urbanelli, S. (2006). Isolation of microsatellite markers in the tiger mosquito *Aedes albopictus* (Skuse). *Mol. Ecol. Notes* 6, 880–881. doi: 10.1111/j.1471-8286.2006.01384.x

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* 155, 945–959.

Pumpuni, C. B., Demaio, J., Kent, M., Davis, J. R., and Beier, J. C. (1996). Bacterial population dynamics in three anopheline species: the impact on Plasmodium sporogonic development. *Am. J. Trop. Med. Hyg.* 54, 214–218.

Raharimalala, F. N., Ravaomanarivo, L. H., Ravelonandro, P., Rafarasoa, L. S., Zouache, K., Tran-Van, V., et al. (2012). Biogeography of the two major arbovirus mosquito vectors, *Aedes aegypti* and *Aedes albopictus* (Diptera, Culicidae), in Madagascar. *Parasit. Vectors* 5, 56. doi: 10.1186/1756-3305-5-56

Ramírez-Puebla, S. T., Rosenblueth, M., Chávez-Moreno, C. K., de Lyra, M. C. P., Tecante, A., and Martínez-Romero, E. (2010). Molecular phylogeny of the genus *Dactylopius* (Hemiptera: Dactylopiidae) and identification of the symbiotic bacteria. *Environ. Entomol.*39, 1178–1183. doi: 10.1603/EN10037

Rani, A., Sharma, A., Rajagopal, R., Adak, T., and Bhatnagar, R. K. (2009). Bacterial diversity analysis of larvae and adult midgut microflora using culture-dependent and culture-independent methods in lab-reared and field-collected *Anopheles stephensi*-an Asian malarial vector. *BMC Microbiol.* 9:96. doi: 10.1186/1471-2180-9-96

R Development Core Team. (2009). *R: A Language and Environment for Statistical Computing.* Vienna: R Foundation for Statistical Computing. Available online at: http://www.R-project.org

Rosenberg, E., and Zilber-Rosenberg, I. (2014). *The Hologenome Concept: Human, Animal and Plant Microbiota*. Berlin Heidelberg: Springer International Publishing.

Rosenberg, N. A. (2004). Distruct: a program for the graphical display of population structure. *Mol. Ecol. Notes* 4, 137–138. doi: 10.1046/j.1471-8286.2003.00566.x

Rueda, L. M. (2004). Pictorial keys for the identification of mosquitoes (Diptera: Culicidae) associated with dengue virus transmission. *Zootaxa* 589, 1–60.

Saboia-Vahia, L., Cuervo, P., Borges-Veloso, A., de Souza, N. P., Britto, C., Dias-Lopes, G., et al. (2014). The midgut of *Aedes albopictus* females expresses active trypsin-like serine peptidases. *Parasit. Vectors* 7:253. doi: 10.1186/1756-3305-7-253

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., et al. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* 75, 7537–7541. doi: 10.1128/AEM.01541-09

Sommer, S. (2005). The importance of immune gene variability (MHC) in evolutionary ecology and conservation. *Front. Zool.* 2:16. doi: 10.1186/1742-9994-2-16

Stouthamer, R., Breeuwer, J. A., and Hurst, G. D. (1999). *Wolbachia pipientis*: microbial manipulator of arthropod reproduction. *Annu. Rev. Microbiol.* 53, 71–102. doi: 10.1146/annurev.micro.53.1.71

Tagliavia, M., Messina, E., Manachini, B., Cappello, S., and Quatrini, P. (2014). The gut microbiota of larvae of *Rhynchophorus ferrugineus* Oliver (Coleoptera: Curculionidae). *BMC Microbiol.* 14:136. doi: 10.1186/1471-2180-14-136

Terenius, O., Lindh, J. M., Eriksson-Gonzales, K., Bussière, L., Laugen, A. T., Bergquist, H., et al. (2012). Midgut bacterial dynamics in *Aedes aegypti*. *FEMS Microbiol. Ecol.* 80, 556–565. doi: 10.1111/j.1574-6941.2012.01317.x

Toft, C., and Andersson, S. G. E. (2010). Evolutionary microbial genomics: insights into bacterial host adaptation. *Nat. Rev. Genet.* 11, 465–475. doi: 10.1038/nrg2798

Tortosa, P., Charlat, S., Labbé, P., Dehecq, J. S., Barré, H., and Weill, M. (2010). *Wolbachia* age-sex-specific density in *Aedes albopictus*: a host evolutionary response to cytoplasmic incompatibility? *PLoS ONE* 5:e9700. doi: 10.1371/journal.pone.0009700

Tortosa, P., Courtiol, A., Moutailler, S., Failloux, A. B., and Weill, M. (2008). Chikungunya-*Wolbachia* interplay in *Aedes albopictus*. *Insect Mol. Biol.* 17, 677–684. doi: 10.1111/j.1365-2583.2008.00842.x

Urbanski, J. M., Benoit, J. B., Michaud, M. R., Denlinger, D. L., and Armbruster, P. (2010). The molecular physiology of increased egg desiccation resistance during diapause in the invasive mosquito, *Aedes albopictus*. *Proc. R. Soc. Lond. B Biol. Sci.* 277, 2683–2692. doi: 10.1098/rspb.2010.0362

Vayssier-Taussat, M., Albina, E., Citti, C., Cosson, J. F., Jacques, M. A., Lebrun, M. H., et al. (2014). Shifting the paradigm from pathogens to pathobiome: new concepts in the light of meta-omics. *Front. Cell. Infect. Microbiol.* 4:29. doi: 10.3389/fcimb.2014.00029

Vaz-Moreira, I., Nunes, O. C., and Manaia, C. M. (2011). Diversity and Antibiotic Resistance Patterns of Sphingomonadaceae Isolates from Drinking Water. *Appl. Environ. Microbiol.* 77, 5697–5706. doi: 10.1128/AEM.00579-11

Vega-Rua, A., Zouache, K., Caro, V., Diancourt, L., Delaunay, P., Grandadam, M., et al. (2013). High efficiency of temperate *Aedes albopictus* to transmit chikungunya and dengue viruses in the Southeast of France. *PLoS ONE* 8:e59716. doi: 10.1371/journal.pone.0059716

Wang, Q., Garrity, G. M., Tiedje, J. M., and Cole, J. R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* 73, 5261–5267. doi: 10.1128/AEM.00062-07

Wang, Y., Gilbreath, T. M., III, Kukutla, P., Yan, G., and Xu, J. (2011). Dynamic gut microbiome across life history of the malaria mosquito *Anopheles gambiae* in Kenya. *PLoS ONE* 6:e24767. doi: 10.1371/journal.pone.0024767

Wang, Y., Wang, Y., Zhang, J., Xu, W., Zhang, J., and Huang, F. S. (2013). Ability of TEP1 in intestinal flora to modulate natural resistance of *Anopheles dirus*. *Exp. Parasitol.* 134, 460–465. doi: 10.1016/j.exppara.2013.04.003

Weiss, B., and Aksoy, S. (2011). Microbiome influences on insect host vector competence. *Trends Parasitol.* 27, 514–522. doi: 10.1016/j.pt.2011.05.001

Yang, Y.-J., Zhang, N., Ji, S.-Q., Lan, X., Shen, Y. L., Li, F. L., et al. (2014). Dysgonomonas macrotermitis sp. *nov.*, isolated from the hindgut of a fungus-growing termite. *Int. J. Syst. Evol. Microbiol.* 64, 2956–2961. doi: 10.1099/ijs.0.061739-0

Ye, L., Amberg, J., Chapman, D., Gaikowski, M., and Liu, W.-T. (2014). Fish gut microbiota analysis differentiates physiology and behavior of invasive Asian carp and indigenous American fish. *ISME J.* 8, 541–551. doi: 10.1038/ismej.2013.181

Zhang, X., Lin, L., Zhu, Z., Yang, X., Wang, Y., and An, Q. (2013). Colonization and modulation of host growth and metal uptake by endophytic bacteria of *Sedum alfredii*. *Int. J. Phytoremediation* 15, 51–64. doi: 10.1080/15226514.2012.670315

Zilber-Rosenberg, I., and Rosenberg, E. (2008). Role of microorganisms in the evolution of animals and plants: the hologenome theory of evolution. *FEMS Microbiol. Rev.* 32, 723–735. doi: 10.1111/j.1574-6976.2008.00123.x

Zouache, K., Fontaine, A., Vega-Rua, A., Mousson, L., Thiberge, J. M., Lourenco-De-Oliveira, R., et al. (2014). Three-way interactions between mosquito population, viral strain and temperature underlying chikungunya virus transmission potential. *Proc. R. Soc. Lond. B Biol. Sci.* 281:20141078. doi: 10.1098/rspb.2014.1078

Zouache, K., Raharimalala, F. N., Raquin, V., Tran-Van, V., Raveloson, L. H. R., Ravelonandro, P., et al. (2011). Bacterial diversity of field-caught mosquitoes, *Aedes albopictus* and *Aedes aegypti*, from different geographic regions of Madagascar. *FEMS Microbiol. Ecol.* 75, 377–389. doi: 10.1111/j.1574-6941.2010.01012.x

Zouache, K., Voronin, D., Tran-Van, V., Mousson, L., Failloux, A. B., Mavingui, et al. (2009). Persistent *Wolbachia* and cultivable bacteria infection in the reproductive and somatic tissues of the mosquito vector *Aedes albopictus*. *PLoS ONE* 4:e6388. doi: 10.1371/journal.pone.0006388

Zurel, D., Benayahu, Y., Or, A., Kovacs, A., and Gophna, U. (2011). Composition and dynamics of the gill microbiota of an invasive Indo-Pacific oyster in the eastern Mediterranean Sea. *Environ. Microbiol.* 13, 1467–1476. doi: 10.1111/j.1462-2920.2011.02448.x

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

# Annexe 2 : Expérience pilote de TD à haut-débit

La mise au point des marqueurs ET par le séquençage à haut-débit a été notamment possible grâce aux enseignements d'une expérience pilote ayant reçu le soutien financier de la Fédération de Recherche 41 "Bio-Environnement Santé".

Ces travaux préliminaires ont été présentés sous la forme d'un poster lors d'une journée organisée par la FR en février 2014.

# Bases génétiques de l'adaptation du moustique tigre *Aedes albopictus* : développement de marqueurs polymorphes

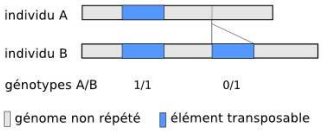
C. Goubert*, H. Henri*, P. Mavingui**, C. Vieira*, G. Minard**, C. Valiente-Moro**[1] et M. Boulesteix*[1]

*UMR 5558 Laboratoire de Biométrie et de Biologie Evolutive | **UMR 5557 Écologie Microbienne | [1]Porteurs du projet devant la FR41
Université Claude Bernard Lyon 1, 43 bd du 11 novembre1918 - 69622 Villeurbanne

## Introduction

Afin de comprendre le succès adaptatif du moustique tigre *Aedes albopictus* en zone tempérée, nous souhaitons comparer le polymorphisme génétique des populations invasives à celui des populations natives d'Asie tropicale, et rechercher par **scan génomique** des régions potentiellement soumises à la sélection.

Pour cela, nous développons une méthode basée sur le **polymorphisme d'insertion des éléments transposables**.




Nous avons testé ce protocole (Transposon Display) sur 4 individus : après amplification spécifique, les insertions ont été séquencées en NGS sur le 454 Junior du DTAMB grâce au financement de la FR41.

**L'objectif de cette expérience pilote était de quantifier ces insertions pour trois familles d'éléments, d'en mesurer le polymorphisme en relation avec la profondeur de séquençage et de tester la spécificité des marqueurs.**

## Matériel et Méthodes

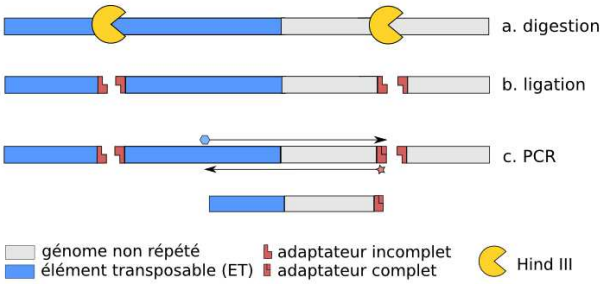



**Figure 1.** Protocole de Transposon Display : amplification spécifique des insertions pour chacune des familles d'ET. **a.** L'ADN génomique est digéré avec HindIII. **b.** Des adaptateurs incomplets sont ligués aux extrémités des fragments de restriction. **c.** PCR : une amorce spécifique de la partie terminale de l'ET permet de compléter l'adaptateur, où une seconde amorce peut alors se fixer et permettre l'amplification de l'insertion

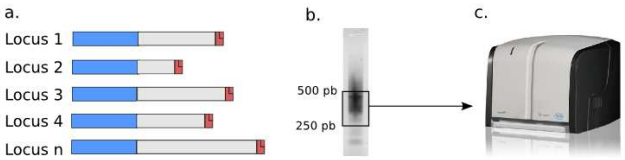



**Figure 2.** Les produits de PCR contenant les différentes insertions **(a.)** sont déposés sur gel d'agarose puis sélectionnés entre 250 et 500 pb **(b.)** avant d'être séquencés par NGS sur 454 Junior. **(c.)**

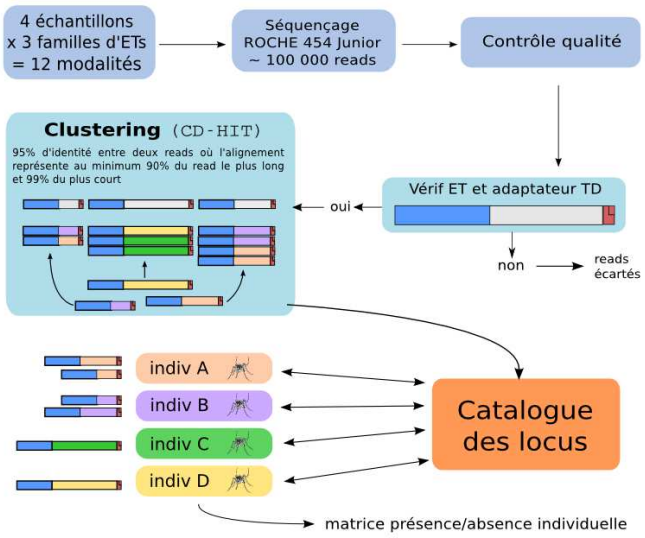



**Figure 3.** Schéma de l'analyse des données : 12 modalités, représentant les insertions de 3 familles d'ET chez 4 individus ont fourni 104 134 reads après leur séquençage. Les contrôles de qualité et de spécificité sont alors effectués avant l'étape de clustering qui permet de regrouper les reads du même locus d'insertion ensemble et de construire un catalogue. Celui-ci permet au final de recenser la présence ou l'absence d'un locus pour chacun des individus.

## Résultats

- **104 134 reads** séquencés, taille moyenne : 271,18 pb

- Un total de **1 091 locus** d'insertion a pu être détecté. Seules deux insertions sont partagées par les 4 individus (monomorphes).

- La majorité de ces locus ont une **profondeur supérieure à 10X**.

- **184** locus sont partagés par au moins 2 individus.

- Nombre moyen de locus par individu : **324** [258 - 421]

| Famille d'ET | Spécificité | # Locus | Profondeur moyenne |
|---|---|---|---|
| **LOA 2B** | 28 % | 219 | 17.2 X |
| **RTE 5** | 43 % | 456 | 16.5 X |
| **I Loner 1** | 89 % | 416 | 15 X |

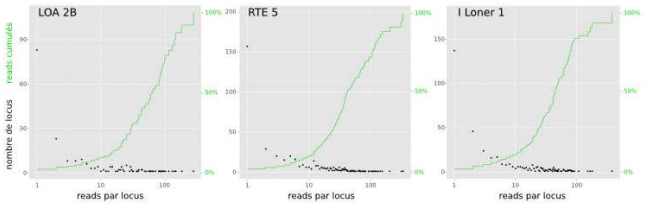



**Figure 4.** Distribution de la profondeur de séquençage des locus d'insertion, par famille d'ET. La fréquence cumulée des reads (courbe verte) indique que leur majorité constitue des locus ayant une profondeur supérieure à 10X.

## Discussion & Perspectives

Cette expérience pilote nous a permis de **valider notre protocole de Transposon Display**, basé sur le séquençage à haut débit.

**Ces résultats encouragent notre stratégie** : à cette profondeur, plus d'un millier de locus hautement polymorphes sont détectés. La spécificité des marqueurs corrobore nos résultats précédents et sera améliorée à l'aide de PCR nichées.

Dans nos applications à venir, l'**augmentation de la profondeur de séquençage** devra permettre d'augmenter le nombre de locus détectés tout en réduisant le nombre de faux négatifs (insertion présente mais non séquencée).